



HAL
open science

Self-Organizing Maps for Exploration of Partially Observed Data and Imputation of Missing Values

Sara Rejeb, Catherine Duveau, Tabea Rebafka

► **To cite this version:**

Sara Rejeb, Catherine Duveau, Tabea Rebafka. Self-Organizing Maps for Exploration of Partially Observed Data and Imputation of Missing Values. *Chemometrics and Intelligent Laboratory Systems*, 2022. hal-03565651v2

HAL Id: hal-03565651

<https://hal.science/hal-03565651v2>

Submitted on 7 Jul 2022 (v2), last revised 10 Feb 2023 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Self-Organizing Maps for Exploration of Partially Observed Data and Imputation of Missing Values

Sara Rejeb^{a,b,*}, Catherine Duveau^b, Tabea Rebafka^a

^a*LPSM, Sorbonne Université, Université de Paris & CNRS, 4, Place Jussieu, 75252, Paris cedex 05, France*

^b*Safran Aircraft Engines, Réau, 77550 Moissy-Cramayel, France*

Abstract

The self-organizing map is an unsupervised neural network which is widely used for data visualisation and clustering in the field of chemometrics. The classical Kohonen algorithm that computes self-organizing maps is suitable only for complete data without any missing values. However, in many applications, partially observed data are the norm. In this paper, we propose an extension of self-organizing maps to incomplete data via a new criterion that also defines estimators of the missing values. In addition, an adaptation of the Kohonen algorithm, named missSOM, is provided to compute these self-organizing maps and impute missing values. An efficient implementation is provided. Numerical experiments on simulated data and a chemical dataset illustrate the short computing time of missSOM and assess its performance regarding various criteria and in comparison to the state of the art.

Keywords: Self-organizing maps, partially observed data, missing data imputation, robustness to missingness mechanism

1. Introduction

In chemometrics, data exploration is essential and is generally the first part of the analysis of any dataset. With the explosion of data volume in many fields of application due to big data, the structure of data is often hidden behind the massive data and difficult to detect. This increases the importance of data exploration methods providing meaningful information on the internal structure and correctness of the data, as well as the relationships and redundancies among the variables. The

*Corresponding author

Email address: sara.rejeb@sorbonne-universite.fr (Sara Rejeb)

initial understanding of the data gained by exploratory data analysis is particularly useful for data modeling.

Common tasks of data exploration are visualization and clustering of the data. While there is plethora of methods addressing one of the tasks, self-organizing maps simultaneously provide both a low-dimensional visual data representation in form of a map and a clustering of the observations. Introduced by [1], this approach consists in mapping the data and performing vector quantization of the input space while preserving topological properties of the data even if these data are high-dimensional. These are good properties that other methods do not have such as k-means which only categorises data without preserving its topology and does not allow data visualisation. Self-organizing maps have become very popular in many applications, since they provide easily interpretable results with a global view of the data. Among others, self-organizing maps are widely used in chemometrics [2, 3, 4, 5, 6, 7], but also in biology [8], humanities [9], industry, such as health monitoring of aircraft engines [10]. Many variants of the standard self-organizing map have been developed, such as Generative Topographic Mapping [11], which is a probabilistic version of the self-organizing map, or extensions to more complex data types (mixed, textual, etc.) [12, 13] demonstrating the relevance of self-organizing maps until today.

A common issue with datasets in most fields of application are missing data. Data may be incomplete for a large variety of reasons. In surveys, for instance, they occur due to non-responses to questions that affect privacy [14, 15, 16]. In industrial applications and chemometrics, measuring instruments may have malfunctions or detection limits yielding erroneous and missing entries [17, 18]. In medical research, missing data can occur in clinical trials when patients abandon or stop taking the treatment for a certain period of time [19, 20, 21]. Missing data also frequently occur in chemical [22, 23] and environmental [24] studies. Moreover, concerning huge databases, merging several datasets from different sources can also result in missing data, as some entries may not be recorded at all for some of the sources.

The impact of missing data on statistical results can be serious, leading to biased estimates, loss of information, decreased statistical power, increased standard errors, and weakened generalizability of findings. However, for a long time, in statistics, missing data have been treated in a very simple and inappropriate way, either by deletion of incomplete measurements or by basic data completion by mean or median values. This has changed during the last decades, by the development of many statistical methods that account for missing data in a meaningful way. There are two general

approaches to deal with missing data: either a statistical method is directly adapted to the partially observed data, or first an appropriate imputation method is applied to complete the data such that the statistical method of interest can be used on the completed or augmented data.

In this paper we are interested in self-organizing maps in the presence of missing data. This problem has been considered among others by [25] by simply restricting all vector calculations to the observed entries. As such, all observed data entries are taken into account in the algorithm. However, the method performs rather poorly when the number of incomplete observations with multiple missing entries is large. Moreover, the imputation of missing data is done afterwards by replacing missing entries by the closest features on the learned map. Related approaches are presented in [26, 27, 28, 29, 30] and more recently in [31, 32].

Missing data imputation can be useful to avoid incomplete data, which is crucial in data mining when methods do not allow for any missing entries. We have the ambition to combine the tasks of imputation and learning the map by a principled approach. Our motivation is the fact that any non trivial imputation method is based on some data model, and so it is natural to use the self-organizing map for imputation. Conversely, a better map may be learned when data are complete. Thus, treating both tasks simultaneously may be beneficial for the two of them.

Our approach can be viewed as an extension of the standard Kohonen algorithm for self-organizing maps and the principle of our method is given in Section 2. A mathematical presentation of the method, a new loss function that encodes our double goal of imputation and learning a self-organizing map and two algorithmic solutions are given in Section 3. Moreover, Section 4 provides an extensive numerical study assessing the robust performance of our method in various settings and in comparison to alternative methods from the literature.

2. The new method in a nutshell

This section presents the principle of our new method for self-organizing maps for partially observed data. To start with, we recall the classical self-organizing map. Let x_1, \dots, x_n be n observations or measurements of dimension p . A self-organizing map represents a nonlinear projection of the high-dimensional data onto a low-dimensional subspace. This subspace is a two-dimensional map represented as a regular grid composed of K fixed neurons. This fixed spatial arrangement of the neurons on the map is the key for the preservation of the topology of the input data when projected onto the map. Every neuron k is associated with a p -dimensional prototype vector w_k ,

also called code vector, that is to be learned. The prototype vectors define a discretization of the data space, and each observation x_i is assigned to its closest prototype. Ideally, prototype vectors of neighboring neurons on the grid are close one to another, so that data points x_i that are close in the input space are also close on the map.

The Kohonen algorithm computes the self-organizing map for complete data in an iterative fashion. One randomly picked observation x_i is treated at each iteration. First, the winning neuron or best matching unit is determined, which is the neuron whose prototype is the closest to measurement x_i . Then, all code vectors w_k are updated by attracting them towards the measurement x_i . The attraction is the strongest for the winning neuron and very weak for the neurons that are far from the winning neuron. Those updates eventually result in an ordered map, where neighboring neurons have similar prototype vectors.

Now, when some of the measurement vectors x_i contain missing entries, the Kohonen algorithm is not applicable anymore. We propose to learn the missing entries while learning the map in the following way. Our algorithm is the following: starting from some initial imputed values, like the mean values computed over the observed entries, select every measurement x_i once, determine its winning neuron by considering the distance only over the observed entries of x_i and then update the code vectors just as in the classical Kohonen algorithm. Then, perform an update of the imputed values using a weighted means of the closest code vectors of the partially observed measurement. Repeat this procedure until convergence. A full description is provided in Algorithm 3.

This algorithm performs both data visualization and imputation of the incomplete data. Interestingly, it is as fast as the standard Kohonen algorithm, since the update of the imputed values is immediate.

3. Self-organizing maps with incomplete data

In this section we first formally state the classical self-organizing map, before introducing the new loss function and two algorithms for the computation of self-organizing maps with partially observed data and missing data imputation.

3.1. Classical Kohonen algorithm

The data matrix containing the measurements is denoted by $X = [x_1, \dots, x_n] \in \mathbb{R}^{n \times p}$. The arrangement of the neurons on the map is given by some neighborhood function $V_\lambda : \{1, \dots, K\}^2 \mapsto$

Algorithm 1: Standard Kohonen algorithm

Input: Data matrix X , size and topology of the map, neighborhood function V_λ , sequence of radii $(\lambda_t)_{0 \leq t \leq T}$ and learning steps $(\varepsilon_t)_{0 \leq t \leq T}$.

Initialize code vectors $W^{(0)}$;

Initialize the counter of iterations: $t = 0$;

while *not converged* **do**

 Increment t : Set $t = t + 1$;

 Choose an observation $i \in \{1, \dots, n\}$ randomly ;

 Assignment: Compute winning neuron $\ell = h(x_i, W^{(t-1)})$;

 Update code vectors:

for $k = 1, \dots, K$ **do**

$w_k^{(t)} = w_k^{(t-1)} + \varepsilon_t V_{\lambda_t}(k, \ell) (x_i - w_k^{(t-1)})$.

end

end

Output: Code vectors $W^{(t)}$.

\mathbb{R}_+ . The neighborhood radius $\lambda > 0$ describes the zone of influence around a neuron. The best prototype vectors of the self-organizing map are defined as the minimum of the loss function F defined by

$$F(W) = \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^K V_\lambda(k, h(x_i, W)) \|x_i - w_k\|_2^2, \quad (1)$$

where $W = [w_1, \dots, w_K] \in \mathbb{R}^{p \times K}$ is the matrix of K prototype vectors and $h : \mathbb{R}^p \times \mathbb{R}^{p \times K} \mapsto \{1, \dots, K\}$ denotes the allocation function, attributing the closest prototype to a data point x w.r.t. the Euclidean distance, defined as

$$h(x, W) = \arg \min_{1 \leq k \leq K} \|x - w_k\|_2. \quad (2)$$

The loss F takes into account all distances between every measurement and all code vectors, weighted by the neighborhood function evaluated on the corresponding neurons. As a result, code vectors that minimize the loss are similar if they are close on the map. In the specific case where the neighborhood function satisfies $V_\lambda(k, \ell) = 0$ for all $k \neq \ell$, the loss F is the criterion minimized by the k -means algorithm. In that case, the clusters are independent unlike SOM where clusters (nodes) are connected to each other.

To compute the minimum of loss F , [33] showed that in the given framework a gradient descent algorithm can be used, referred to as the Kohonen stochastic algorithm. For a randomly picked observation x_i with winning neuron $\ell = h(x_i, W^{(t)})$, the updates of the code vectors are given by

$$w_k^{(t+1)} = w_k^{(t)} + \varepsilon_t V_\lambda(k, \ell)(x_i - w_k^{(t)}), \quad (3)$$

where $(\varepsilon_t)_{t \geq 0}$ is a sequence of decreasing learning steps. This update attracts all prototypes towards observation x_i . It is also common to shrink the neighborhood by using a decreasing sequence of radii $(\lambda_t)_{t \geq 0}$ in the neighborhood function V_λ . The algorithm is summarized in Algorithm 1.

3.2. Notation for incomplete data

Now we consider an incomplete $n \times p$ data matrix containing missing values. Let the matrix $M = (m_{i,j})_{i,j} \in \{0, 1\}^{n \times p}$ be the missing-data pattern which indicates where the entries are missing or masked, and that is defined by

$$m_{i,j} = \begin{cases} 1 & \text{if } x_{i,j} \text{ is observed} \\ 0 & \text{if } x_{i,j} \text{ is missing} \end{cases}$$

We denote X^{obs} the set of observed data values and X^{miss} the set of non-observed data entries hidden by the missing-data pattern M . The complete data are denoted by $X^{\text{compl}} = (X^{\text{obs}}, X^{\text{miss}})$. Likewise, for the observation vector x_i we denote by x_i^{obs} and x_i^{miss} the observed and unobserved entries, respectively, and, with some abuse of notation, $x_i^{\text{compl}} = (x_i^{\text{obs}}, x_i^{\text{miss}})$ is the complete vector, which also corresponds to the i -th row of X^{compl} .

Our goal is to adapt the model of self-organizing maps to partially observed data, and moreover, learn the values of the missing data. The motivation to treat these tasks simultaneously is that learning missing values requires a data model, and as we are interested in self-organizing maps it is natural to use this model for data imputation. At the same time, learning a map with completed data may give better results compared to using only the observed part X^{obs} of the data.

3.3. New loss function

We introduce a new loss function that considers both problems : finding the best self-organizing map and the best values for imputation of the missing data. In other words, by minimizing the new loss function F_{missom} we search for both the best code vectors $W \in \mathbb{R}^{p \times K}$ for the map and the best

values for the missing data denoted by X^* chosen in the set of all possible values for the missing entries $\mathcal{X}^{\text{miss}}$.

To define the new criterion, an adaptation of the definition of the winning neuron is in order. In the presence of missing values, it is natural to restrict the Euclidean distance in (2) only to the observed entries. More precisely, for any vectors $x^{\text{obs}} \in \mathbb{R}^{p'}$ ($p' \leq p$), $m \in \{0, 1\}^p$ with $\sum_{j=1}^p m_j = p'$ and code vectors $W \in \mathbb{R}^{p \times K}$, we set

$$h^{\text{miss}}(x^{\text{obs}}, m, W) = \arg \min_{1 \leq k \leq K} \|x^{\text{obs}} - w_k \odot m\|_2,$$

where $w_k \odot m$ denotes the p' -vector made of the elements $w_{k,j}$ of w_k such that $m_j = 1$. Now, we define the new loss as

$$F_{\text{missom}}(W, X^*) = \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^K V_\lambda(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W)) \|(x_i^{\text{obs}}, x_i^*) - w_k\|_2^2,$$

where $m_i \in \mathbb{R}^p$ is the i -th row of the matrix M and $(x_i^{\text{obs}}, x_i^*)$ denotes the i -th measurement vector completed with x_i^* . Since

$$\|(x_i^{\text{obs}}, x_i^*) - w_k\|_2^2 = \|x_i^{\text{obs}} - w_k \odot m_i\|_2^2 + \|x_i^* - w_k \odot (\mathbf{1}_p - m_i)\|_2^2,$$

where $\mathbf{1}_p = (1, \dots, 1)^T \in \mathbb{R}^p$, the criterion F_{missom} can be decomposed into two parts according to the observed and the missing entries as

$$F_{\text{missom}}(W, X^*) = F_{\text{obs}}(W) + F_{\text{miss}}(W, X^*),$$

where $F_{\text{obs}}(W)$ is the part of the loss over the observed entries given by

$$F_{\text{obs}}(W) = \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^K V_\lambda(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W)) \|x_i^{\text{obs}} - w_k \odot m_i\|_2^2,$$

and $F_{\text{miss}}(W, X^*)$ is the contribution of the imputed values X^* to the loss, defined as

$$F_{\text{miss}}(W, X^*) = \frac{1}{2n} \sum_{i=1}^n \sum_{k=1}^K V_\lambda(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W)) \|x_i^* - w_k \odot (\mathbf{1}_p - m_i)\|_2^2.$$

Note that in the complete-data case, where the missing-data pattern is $M = \mathbf{1}_{n \times p}$, $F_{\text{miss}}(W, X^*) = 0$ for any W and any X^* , so that the criterion F_{missom} is equal to the one of the classical self-organizing map, that is, $F_{\text{missom}}(W, X^*) = F(W)$.

Algorithm 2: missSOM algorithm

Input: Incomplete data X^{obs} , missing-data pattern M , size and topology of the map, neighborhood function V_λ , sequence of radii $(\lambda_t)_{0 \leq t \leq T}$ and learning steps $(\varepsilon_t)_{0 \leq t \leq T}$.

Initialize imputed values $X^{*(0)}$ and code vectors $W^{(0)}$;

Initialize the counter of iterations: $s = 0$;

while *not converged* **do**

 Increment s : Set $s = s + 1$;

 Update code vectors by Kohonen Algorithm 1 on the augmented data with winning neurons obtained by h^{miss} instead of h :

$W^{(s)} \leftarrow \text{Kohonen}(X^{\text{aug}} = (X^{\text{obs}}, X^{*(s-1)}))$;

 Update imputed values: for i, j such that $m_{i,j} = 0$,

$$x_{i,j}^{*(s)} = \frac{\sum_{k=1}^K V_{\lambda_T}(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W^{(s)})) w_{k,j}^{(s)}}{\sum_{k=1}^K V_{\lambda_T}(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W^{(s)}))}.$$

end

Output: Code vectors $W^{(s)}$ and imputed data $X^{*(s)}$.

3.4. Minimization algorithm

For the minimization of $(W, X^*) \mapsto F_{\text{miss}}(W, X^*)$ on $\mathbb{R}^{K \times p} \times \mathcal{X}^{\text{miss}}$ we propose to alternate the minimization in W and X^* while keeping the other argument fixed.

For fixed X^* , the function $W \mapsto F_{\text{missom}}(W, X^*)$ is similar to the objective function F in (1) in the complete-data case applied to the augmented data $X^{\text{aug}} = (X^{\text{obs}}, X^*)$. The only difference lies in the definition of the winning neurons by h^{miss} that appear in the neighbourhood function V_λ . Thus, a Kohonen algorithm applied to X^{aug} can be used to find the best code vectors W .

In turn, when W is fixed, the minimization of $X^* \mapsto F_{\text{missom}}(W, X^*)$ boils down to minimize $X^* \mapsto F_{\text{miss}}(W, X^*)$. This problem has a unique explicit solution given for all $1 \leq i \leq n$ and j such that $m_{i,j} = 0$ by

$$x_{i,j}^* = \frac{\sum_{k=1}^K V_\lambda(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W)) w_{k,j}}{\sum_{k=1}^K V_\lambda(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W))}. \quad (4)$$

That is, the imputed values are a weighted mean of the prototype vectors weighted according to the neighborhood function.

To summarize, the algorithm updates imputed values for the missing data and applies the classical Kohonen algorithm with adjusted winning neuron function h^{miss} to learn the map. This is repeated until convergence or until a maximum number of iterations chosen by the user is attained. The algorithm is described in Algorithm 2.

As initial values for the imputed values X^* , one can simply impute the sample mean or median of the variables obtained over the observed entries.

3.5. Accelerated version

Algorithm 2 happens to be expensive in terms of computing time, in particular when the number of iterations is large, since the entire standard Kohonen algorithm is carried out during each iteration. A speed up is obtained by intertwining updates of the missing data and the iterations of the Kohonen Algorithm 1. More precisely, we propose to update the missing data at every epoch, that is, after every pass through the data. This procedure gives rise to Algorithm 3. As such, the Kohonen algorithm is carried out only once, while in the initial Algorithm 2, the entire Kohonen algorithm is applied repeatedly. Thus the computing time of the accelerated version of missSOM is comparable to the computing time of the standard Kohonen Algorithm 1, since the update of the imputed values is fast.

While the first version of the missSOM algorithm has some theoretical justification, the accelerated version lacks this foundation. A numerical study given in Appendix 6.1 shows that the Algorithms 2 and 3 provide very similar maps and hence justifies the utilization of the accelerated version, which achieves a significant gain in computing time.

Note that while the selection of the observations x_i in Algorithm 3 is deterministic, it is possible to use a random selection scheme.

4. Numerical experiments

In this section, the performance of the proposed method missSOM is evaluated and compared to alternative methods. A simulation study is conducted to assess the quality of the representation of the data via the map and the accuracy of imputed values under various conditions.

4.1. Performance criteria

Let W^* be the code vectors of the final self-organizing map, X^* the imputed values and $\hat{x}_i = (x_i^{\text{obs}}, x_i^*)$ the completed observation vectors.

The quality of the map as a representation of the data can be evaluated by two criteria. First, the *quantization error* defined as the average of the squared distances between the observations and their nearest prototype vector given by

$$E = \frac{1}{n_{\text{obs}}} \sum_{i=1}^n \|x_i^{\text{obs}} - w_{h^{\text{miss}}(x_i^{\text{obs}}, m_i, W^*)}\|_2^2,$$

where $n_{\text{obs}} = \sum_{i,j} m_{i,j}$ is the number of observed entries in the data, informs on whether the prototype vectors are good representations of the data. Second, the *topographic error* evaluates the preservation of the topology of the data in the map by the proportion of observations for which the winning neuron and the second closest neuron are not neighbors, i.e. not connected on the grid. It is defined as

$$T = \frac{1}{n} \sum_{i=1}^n e(\hat{x}_i),$$

where

$$e(\hat{x}_i) = \begin{cases} 0 & \text{if the neurons } h^{\text{miss}}(x_i^{\text{obs}}, m_i, W^*) \text{ and } v^{\text{miss}}(x_i^{\text{obs}}, m_i, W^*) \text{ are adjacent} \\ 1 & \text{otherwise} \end{cases}$$

with $v^{\text{miss}}(x_i^{\text{obs}}, m_i, W^*) = \arg \min_{j \neq h^{\text{miss}}(x_i^{\text{obs}}, m_i, W^*)} \|x_i^{\text{obs}} - w_k^* \odot m_i\|_2^2$ the second closest neuron to measurement x_i^{obs} . The topographic error is expected to be small when the map is well organized and ordered. As both errors are computed only on the observed entries, the focus is put on the quality of the map with respect to the observed part of the data. The accuracy of the imputed values is assessed separately by the *imputation error*, which is defined as the root mean square error and quantifies the quality of the imputed values compared to the true missing values.

4.2. Datasets

For the numerical experiments, two settings are considered. First, the dataset *wines* from the UCI machine learning repository [34] is used, which contains the results of a chemical analysis of 178 wines on 13 quantitative variables. We generate 100 perturbed datasets by adding gaussian noise with mean 0 and standard deviation equal to one tenth of the mean value in each variable.

In the second setting, 100 datasets are drawn from a multivariate gaussian mixture with dimension $p = 5$, four equal-sized groups and $n = 2000$ observations. The correlation among all pairs of components is equal to 0.5 and for every dataset, the 4 gaussian means are drawn independently from a centered normal distribution with standard deviation equal to 5.

In both settings, missing values are generated using the `ampute` function from the R package `mice` [35]. Different proportions of missing values, namely 5%, 20% and 40%, and different mechanisms of missingness are considered. The literature on missing data traditionally distinguishes three mechanisms that lead to missingness [36] and it is well known that the performance of imputation methods may be sensitive to the mechanism at work. The mechanism is said to be *missing completely at random* (MCAR), when the causes of missing values are independent from the data and the probability of being absent is the same for all items. That is, a subset of observations is chosen at random using independent Bernoulli variables with fixed success probability for all entries. In contrast, when the probability that a value is missing depends on the values of the observed variables, the mechanism is called *missing at random* (MAR). In our simulations, to obtain MAR, for each variable, missing values are obtained using a logistic regression model depending only on the other variables. Finally, when the probability of being absent also depends on the unobserved value, the mechanism is called *missing not at random* (MNAR) and we can consider missing data simulated by using a logistic regression model depending on all variables.

4.3. Alternative methods

Our study is two fold. On the one hand, we compare *missSOM* to the state of the art on self-organizing maps. The simplest method (referred to as *deletion* in the tables) consists in deleting the observations containing missing values and applying the standard Kohonen algorithm to the remaining data. Incomplete observations are classified once the map is built by assigning them to their closest prototypes and missing entries are imputed by the corresponding values of the winning prototypes. *Cottrell's* approach is a variant of the classical self-organizing maps appropriate to deal with incomplete observations [25]. *Cottrell* adapts the Kohonen algorithm by simply restricting all vector calculations to the observed entries. The main difference with *missSOM* is that data imputation is performed only after learning the map by imputation with the values of the closest prototypes.

On the other hand, in our experiments, we compare *missSOM* to the state of the art on missing data imputation. The literature provides numerous general imputation methods. A basic approach (here referred to as *mean*) is the imputation by the mean value of the observed variables. A non-parametric approach called *missForest* predicts missing values using a random forest trained on the observed parts of the dataset. Moreover, the method referred to as *knn* is a k -nearest neighbors

approach, which is implemented in the `VIM` package [37]. Missing values are imputed iteratively by a weighted average of the k closest observations. Finally, assuming a gaussian mixture model for the data, the `Amelia` package [38] performs imputation using the expectation-maximization algorithm and a bootstrap approach to iteratively estimate missing values. In our numerical study, the imputation error of `missSOM` is compared to the one obtained by all these methods. In addition, we also compare the map obtained by `missSOM` with the maps obtained by the classical Kohonen algorithm applied to the complete dataset, where missing values are imputed by the aforementioned imputation methods.

As a benchmark we consider SOM trained on the complete data set (corresponding to a missing rate of 0%) and compute the error rates only on the observed values that are provided to the other methods. This method is referred to as *complete-data SOM*.

4.4. Parameters of the methods

The `missSOM` algorithm is applied with the default parameters in the package `missSOM`. That is, the neighborhood function is a gaussian and the sequence $(\lambda_t)_{t \leq T}$ decreases from $\lambda_0 = 4.58$ to $\lambda_T = 0.5$ for the wine dataset and from $\lambda_0 = 8.89$ to $\lambda_T = 0.5$ for the gaussian mixture data. The maximum number of epochs is $T = 100$. Concerning the map, it has a hexagonal topology and the grid is composed of $K = 9 \times 7$ neurons for the wines data and $K = 16 \times 14$ neurons for the gaussian mixture data. The choice of the size of the map depends on the sample size and on the objective of the analysis. In general, `missSOM` has the same parameters as classical SOM and they can be chosen in the same way as for SOM.

For the alternative methods, the maximum number of iterations is set to 100 and for the other parameters the default values are used.

4.5. Comparison to the state of the art on self-organizing maps

Figure 1 and 2 present the boxplots of the topographic and the quantization error for all SOM-type methods in both settings. They serve to evaluate the quality of the representation of the data using a self-organizing map computed with the different approaches. The corresponding imputation errors in Figure 3 allow to judge their performances as imputation methods.

First, we observe that when the percentage of missing data is low, all methods mostly have very similar performances. With increasing percentage of missingness, the problem is harder and

differences among the methods appear. But as errors are evaluated only on the observed part of the data, errors are not necessarily increasing. Concerning the mechanisms of missingness, it appears that the quality of the map does not depend on it for any SOM-type method. However, imputation is impacted by the mechanism. Imputation is the easiest under MCAR and the hardest under MNAR.

Next, we see that the *deletion* method is very unstable. In some scenarios its error rates are among the worst, and, more importantly, when too many values are missing, the method breaks down and does not produce any result. Indeed, on small datasets as wines with 20% of missingness, the number of complete observations is smaller than the size of the map and thus classical SOM is just not applicable. On the large gaussian mixture dataset, the *deletion* method produces the best topographic error, but the associated quantization error is disastrous, disqualifying the approach.

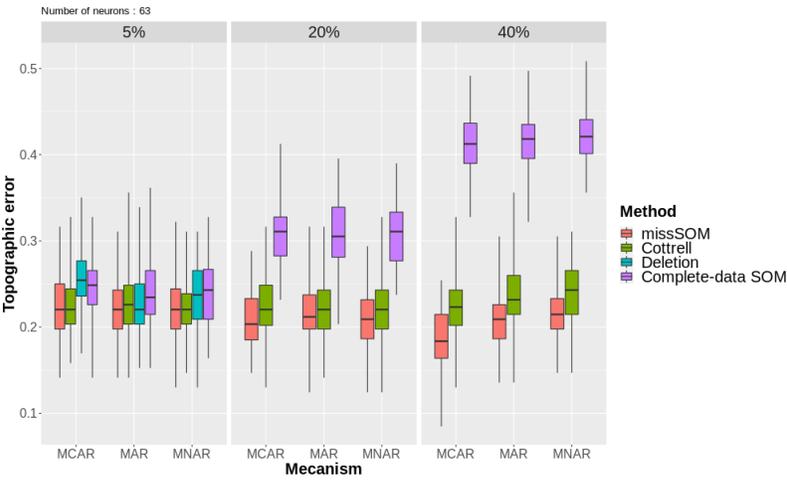
The quantization error of the *complete-data SOM* method is constant when varying the amount of missingness, because the underlying map remains the same. For the topographic error, an increase is observed which is due to the definition of the error. The error determines the closest and second closest prototypes only with respect of the observed entries, while the map was optimized by taking into account the complete data. This explains why *missSOM* and *Cottrell* have lower topographic errors, as their maps are learned with a notion of closeness restricted on the observed entries.

Finally, we observe that *Cottrell's* method always achieves the best quantization errors, directly followed by *missSOM*. Concerning the topographic error, *missSOM* is consistently doing better than *Cottrell*. Thus, in terms of quality of the map and representation of the data, none of the methods outperforms all others, and *Cottrell* and *missSOM* have both a good global performance.

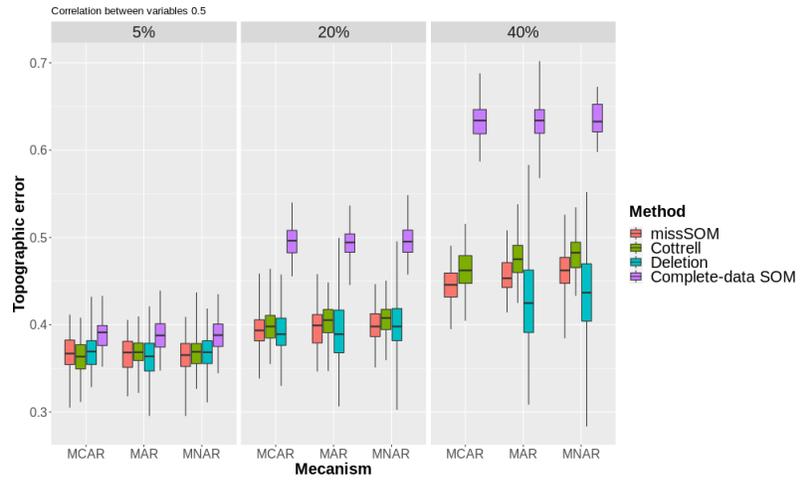
Now, considering the imputation error, there is a clear winner. When the proportion of missingness is important, *missSOM* outperforms all other methods in every scenario. This confirms us in the use of *missSOM* with respect to *Cottrell*, especially when accurate imputation is desired.

4.6. Comparison to the state of the art on missing-data imputation

In the second part of the simulation study, *missSOM* is compared to general imputation methods. Figure 4 shows that as an imputation method *missForest* is unbeaten regardless of the missingness mechanism and the percentage of missing data. On gaussian mixture data, *knn* also provides accurate imputations. The imputation accuracy achieved by *missSOM* is high and *missSOM* is mostly among the best imputation methods.

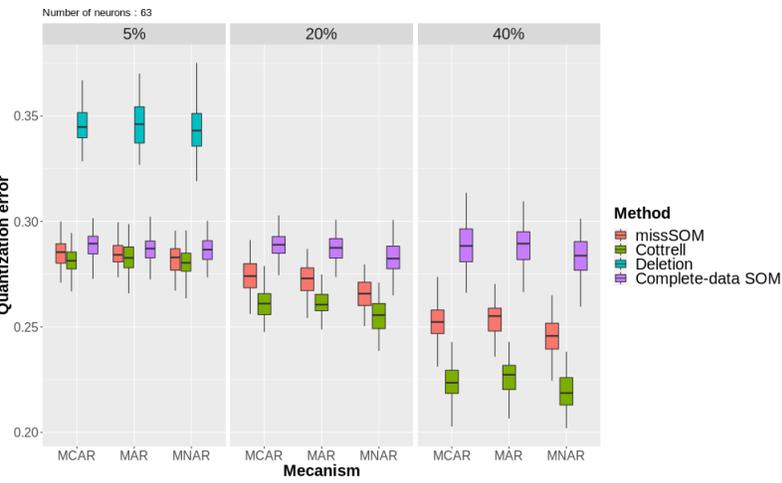


(a) Wines data

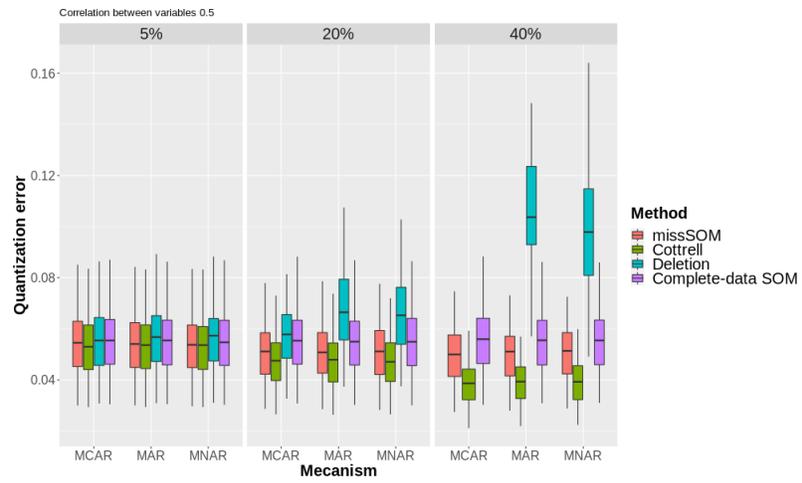


(b) Gaussian mixture data

Figure 1: Topographic error of SOM-type methods for various amounts and mechanisms of missingness on the wines (a) and the gaussian mixture data (b).

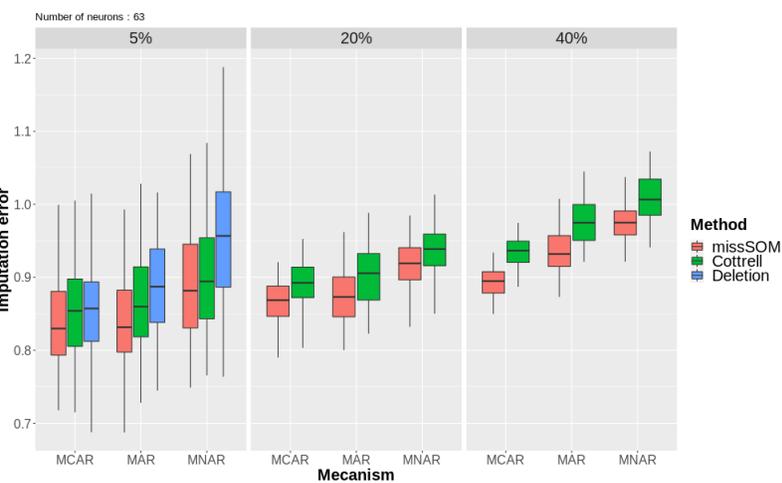


(a) Wines data

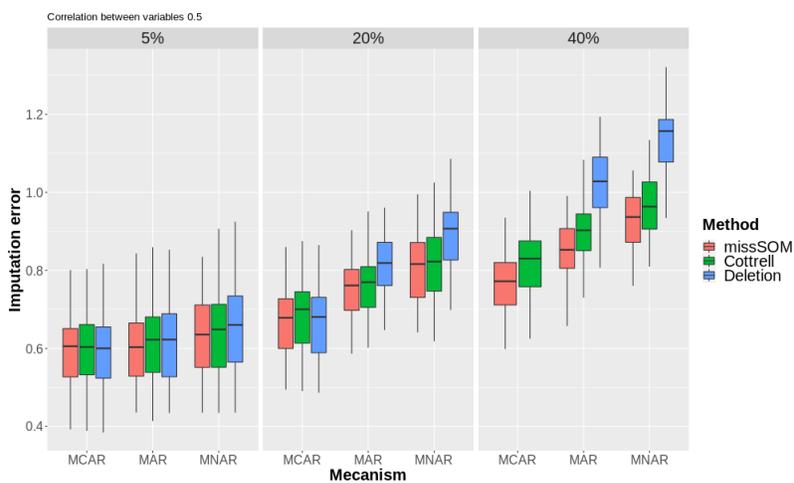


(b) Gaussian mixture data

Figure 2: Quantization error of SOM-type methods for various amounts and mechanisms of missingness on the wines (a) and the gaussian mixture data (b).

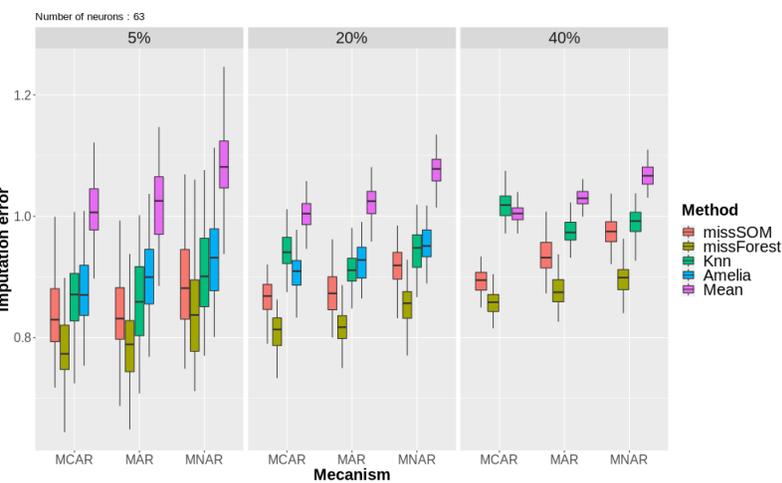


(a) Wines data

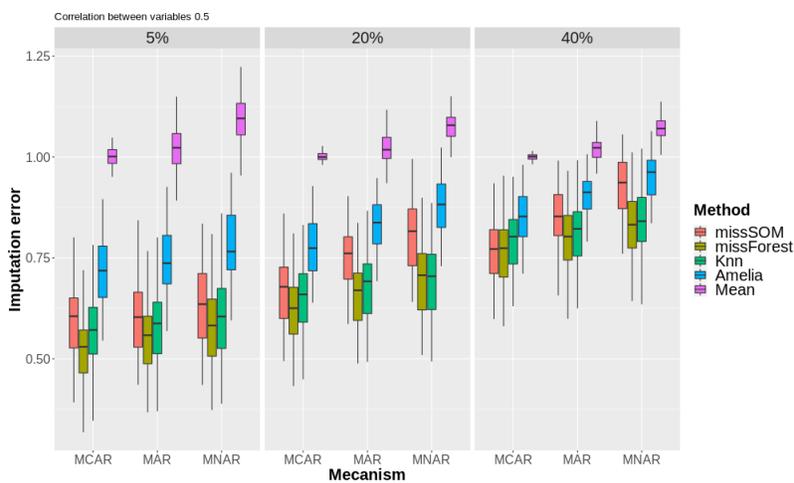


(b) Gaussian mixture data

Figure 3: Imputation error of SOM-type methods for various amounts and mechanisms of missingness on the wines (a) and the gaussian mixture data (b).

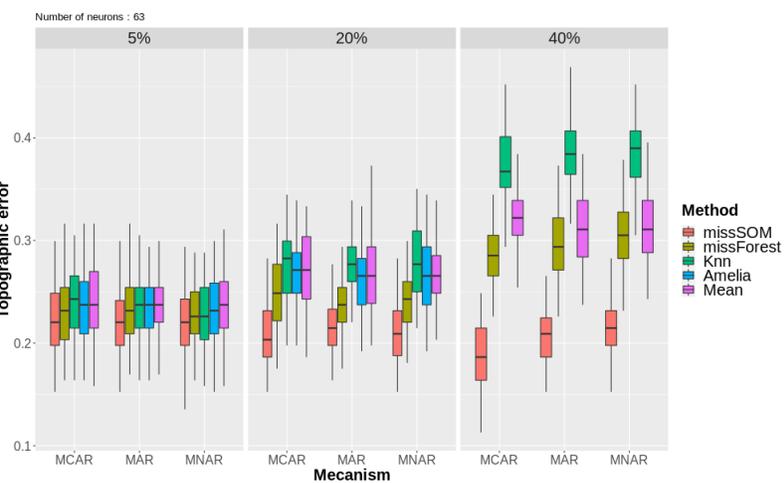


(a) Wines data

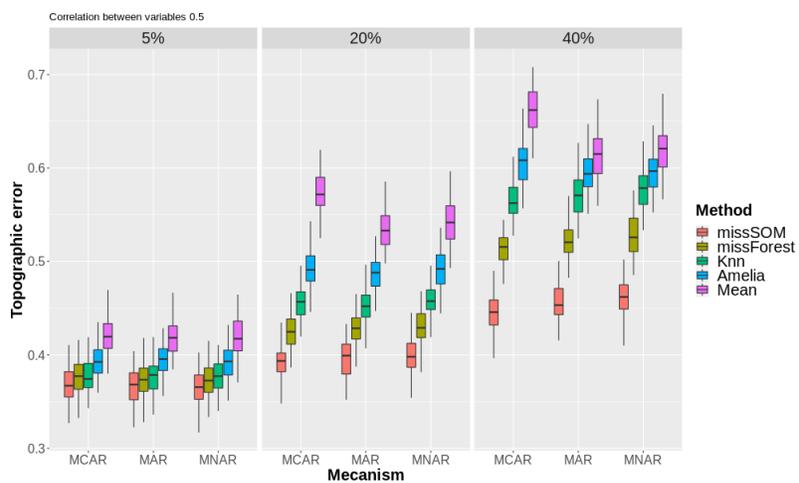


(b) Gaussian mixture data

Figure 4: Imputation error of *missSOM* and classical imputation methods for various amounts and mechanisms of missingness on the wines (a) and the gaussian mixture data (b).

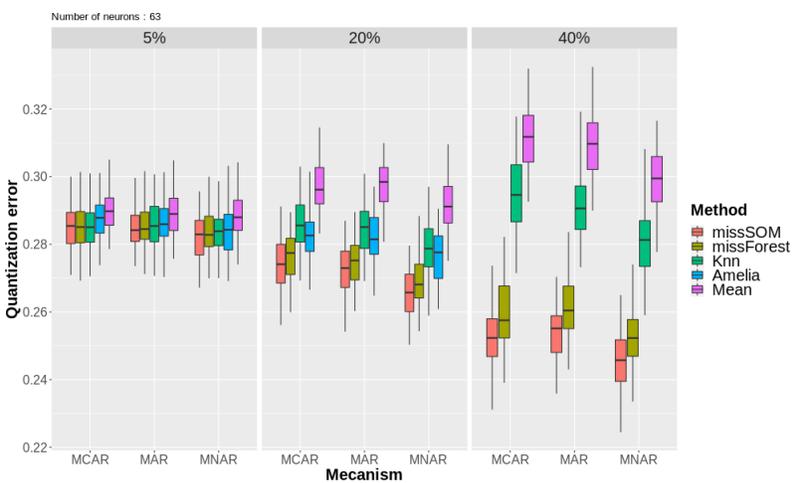


(a) Wines data

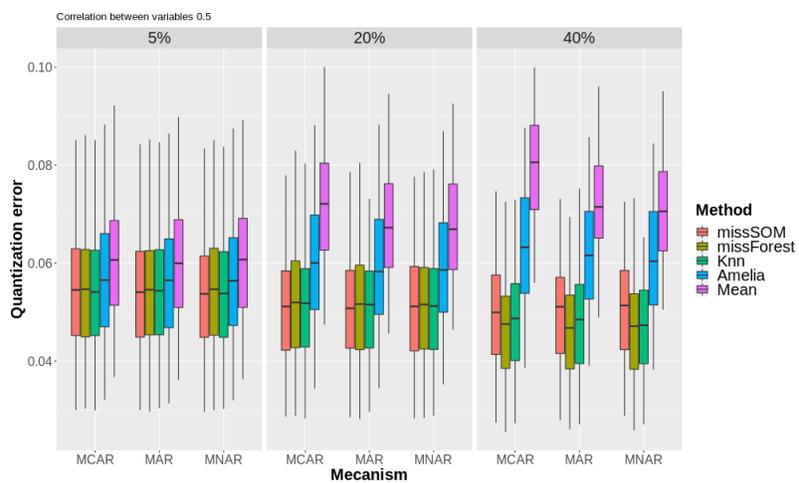


(b) Gaussian mixture data

Figure 5: Topographic error of *missSOM* and classical imputation methods for various amounts and mechanisms of missingness on the wines (a) and the gaussian mixture data (b).



(a) Wines data



(b) Gaussian mixture data

Figure 6: Quantization error of *missSOM* and classical imputation methods for various amounts and mechanisms of missingness on the wines (a) and the gaussian mixture data (b).

Finally, Figure 5 and Figure 6 provide insights on the quality of the data representation, when classical SOM is applied to the data with missing values imputed by the different imputation methods. In this respect, *missSOM* clearly outperforms all the other methods. In particular, *missForest* provides self-organizing maps of poorer quality, especially in terms of preservation of the topology of the input data. This confirms the initial motivation of *missSOM*: in the presence of missing values and when a self-organizing map is required, it is better to learn the map simultaneously with the imputed values than to treat the two tasks separately.

A further advantage of *missSOM* is its computing time, which is about 10 times shorter than the one of *missForest*.

To summarize, we have seen that *missSOM* is the method of choice when both tasks data imputation and data representation are desired. However, when the focus is only on data imputation, then better alternatives as *missForest* exist.

5. Conclusion

In this paper, we have proposed an extension of the self-organizing map for partially observed data, referred to as *missSOM*. The proposed method addresses simultaneously the two problems of computing a self-organizing map and imputing missing values. A numerical study assesses the good performance of *missSOM* regarding various criteria and in comparison to the state of the art. While this paper focuses on the standard Kohonen algorithm, in future work we may address the transfer of our approach to other existing variants of self-organizing maps on more complex data types such as mixed data to enable them to deal with missing data.

6. Appendix

6.1. Validation of accelerated *missSOM* algorithm

Table 1 compares the results of the basic *missSOM* Algorithm 2 and its accelerated version Algorithm 3 on the simulated gaussian mixture data in various conditions. On the one hand, for all settings the errors are totally equivalent. This indicates that the accelerated version provides the same self-organizing maps and very similar imputations as the basic *missSOM* algorithm. On the other hand, we see that in terms of computing time we gain two orders of magnitude. Hence, the use of the accelerated algorithm instead of the basic version is completely justified.

References

- [1] T. Kohonen, *Self-organizing maps*, Springer, Berlin, 1995.
- [2] R. High, G. T. Eyres, P. Bremer, B. Kebede, Characterization of blue cheese volatiles using fingerprinting, self-organizing maps, and entropy-based feature selection, *Food Chemistry* 347 (2021) 128955.
- [3] S. Licen, S. Cozzutto, G. Barbieri, M. Crosera, G. Adami, P. Barbieri, Characterization of variability of air particulate matter size profiles recorded by optical particle counters near a complex emissive source by use of self-organizing map algorithm, *Chemometrics and Intelligent Laboratory Systems* 190 (2019) 48–54.
- [4] T. Voyslavov, S. Tsakovski, V. Simeonov, Surface water quality assessment using self-organizing maps and hasse diagram technique, *Chemometrics and Intelligent Laboratory Systems* 118 (2012) 280–286.
- [5] C. Krongchai, S. Funsueb, J. Jakmune, S. Kittiwachana, Application of multiple self-organizing maps for classification of soil samples in thailand according to their geographic origins, *Journal of Chemometrics* 31 (2) (2017) e2871.
- [6] V. Besada, C. Quelle, J. M. Andrade, N. Gutiérrez, M. P. Gómez-Carracedo, F. Schultze, A 10-year survey of trace metals in sediments using self-organizing maps, *Journal of Chemometrics* 28(7).
- [7] H. Cremasco, D. Borsato, K. G. Angilelli, O. F. Galão, E. Bona, M. E. Valle, Application of self-organising maps towards segmentation of soybean samples by determination of inorganic compounds content, *Journal of the Science of Food and Agriculture* 96 (1) (2016) 306–310.
- [8] P. Melin, J. C. Monica, D. Sanchez, O. Castillo, Analysis of spatial spread relationships of coronavirus (covid-19) pandemic in the world using self organizing maps, *Chaos, Solitons & Fractals* 138 (2020) 109917.
- [9] S. Massoni, M. Olteanu, P. Rousset, Career-Path Analysis Using Optimal Matching and Self-Organizing Maps, in: R. M. José C. Principe (Ed.), *Advances in Self-Organizing Maps*, Lecture Notes in Computer Science n°5629, Springer, 2009, pp. 154–162.

- [10] E. Côme, M. Cottrell, M. Verleysen, J. Lacaille, Aircraft engine health monitoring using Self-Organizing Maps, in: 10th Industrial Conference ICDM, 2010, pp. 405–417.
- [11] C. M. Bishop, M. Svensén, C. K. I. Williams, GTM: The Generative Topographic Mapping, *Neural Computation* 10 (1) (1998) 215–234.
- [12] T. Kohonen, P. Somervuo, How to make large self-organizing maps for nonvectorial data, *Neural Networks* 15 (8) (2002) 945–952.
- [13] M. Lebbah, A. Chazottes, F. Badran, S. Thiria, Mixed Topological Map, *European Symposium on Artificial Neural Networks (ESANN)* (2005).
- [14] E. R. Ziegel, Book review: Nonresponse in household interview surveys by Robert M. Groves; Mick P. Couper, *Technometrics* 41 (4) (1999) 381–381.
- [15] A. G. Blom, E. D. de Leeuw, J. J. Hox, Interviewer effects on nonresponse in the European Social Survey, *ISER Working Paper Series 2010-25*, Institute for Social and Economic Research (2010).
- [16] A. Mirzaei, S. R. Carter, A. E. Patanwala, C. R. Schneider, Missing data in surveys: Key concepts, approaches, and applications, *Research in Social and Administrative Pharmacy* 18 (2021) 2308–2316.
- [17] K. Lakshminarayan, S. Harp, T. Samad, Imputation of missing data in industrial databases, *Applied Intelligence* 11 (2004) 259–275.
- [18] L. Ehrlinger, T. Grubinger, B. Varga, M. Pichler, T. Natschläger, J. Zeindl, Treating missing data in industrial data analytics, in: 2018 Thirteenth International Conference on Digital Information Management (ICDIM), 2018, pp. 148–155.
- [19] W. Shih, Problems in dealing with missing data and informative censoring in clinical trials, *Current Controlled Trials in Cardiovascular Medicine* 3 (2002) 4.
- [20] R. J. Little, R. D’Agostino, M. L. Cohen, K. Dickersin, S. S. Emerson, J. T. Farrar, C. Frangakis, J. W. Hogan, G. Molenberghs, S. A. Murphy, J. D. Neaton, A. Rotnitzky, D. Scharfstein, W. J. Shih, J. P. Siegel, H. Stern, The prevention and treatment of missing data in clinical trials, *New England Journal of Medicine* 367 (14) (2012) 1355–1360.

- [21] G. Campbell, G. Pennello, L. Yue, Missing data in the regulation of medical devices, *Journal of Biopharmaceutical Statistics* 21 (2) (2011) 180–195.
- [22] C. Lazar, L. Gatto, M. Ferro, C. Bruley, T. Burger, Accounting for the multiple natures of missing values in label-free quantitative proteomics data sets to compare imputation strategies, *Journal of Proteome Research* 15 (4) (2016) 1116–1125.
- [23] J. N. Cape, R. I. Smith, D. Leaver, Missing data in spatiotemporal datasets: the UK rainfall chemistry network, *Geoscience Data Journal* 2 (1) (2015) 25–30.
- [24] A. Smolinski, S. Hlawiczka, Chemometric treatment of missing elements in air quality data sets, *Polish Journal of Environmental Studies* 16 (4) (2007) 613–622.
- [25] M. Cottrell, L. P., Missing values : processing with the Kohonen algorithm, in: *ASMDA, 2005*, pp. 489–496.
- [26] L. Folguera, J. Zupan, D. Cicerone, J. F. Magallanes, Self-organizing maps for imputation of missing data in incomplete data matrices, *Chemometrics and Intelligent Laboratory Systems* 143 (2015) 146–151.
- [27] R. Rustum, A. J. Adeloye, Replacing outliers and missing values from activated sludge data using Kohonen Self-Organizing Map, *Journal of Environmental Engineering* 133 (9) (2007) 909–916.
- [28] E. Nkiaka, N. Nawaz, J. Lovett, Using Self-Organizing Maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin, *Environmental Monitoring and Assessment* 188 (7) (2016) 400.
- [29] A. J. Adeloye, R. Rustum, I. D. Kariyama, Neural computing modeling of the reference crop evapotranspiration, *Environmental Modelling & Software* 29 (1) (2012) 61–73.
- [30] A. M. Kalteh, P. Hjorth, Imputation of missing values in a precipitation–runoff process database, *Hydrology Research* 40 (4) (2009) 420–432.
- [31] N. Junno, E. Koivisto, I. Kukkonen, A. Malehmir, M. Montonen, Predicting missing seismic velocity values using self-organizing maps to aid the interpretation of seismic reflection data from the Kevitsa Ni-Cu-PGE deposit in northern Finland, *Minerals* 9 (9) (2019) 529.

- [32] B. Khotimah, Miswanto, H. Suprajitno, A hybrid self organizing map imputation (SOMI) with naïve bayes for imputation missing data classification, *International Journal of GEOMATE* 17 (62) (2019) 195–202.
- [33] H. Ritter, T. Martinetz, K. Schulten, *Neural Computation and Self-Organizing Maps: An Introduction*, Addison-Wesley Longman Publishing Co., Inc., USA, 1992.
- [34] D. Dua, C. Graff, *UCI machine learning repository* (2017).
- [35] S. van Buuren, K. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R, *Journal of Statistical Software, Articles* 45 (3) (2011) 1–67.
- [36] D. B. Rubin, Inference and missing data, *Biometrika* 63 (3) (1976) 581–592.
- [37] A. Kowarik, M. Templ, Imputation with the R package VIM, *Journal of Statistical Software* 74 (7) (2016) 1–16.
- [38] J. Honaker, G. King, M. Blackwell, Amelia II: A program for missing data, *Journal of Statistical Software* 45 (7) (2011) 1–47.

Algorithm 3: Accelerated missSOM algorithm

Input: Incomplete data matrix X^{obs} , missing-data pattern M , size and topology of the map, neighborhood function V_λ , sequence of radii $(\lambda_t)_{0 \leq t \leq T}$ and learning steps

$$(\varepsilon_t)_{0 \leq t \leq T}.$$

Initialize imputed values $X^{*(0)}$ and code vectors $W^{(0)}$;

Initialize the number of epochs: $t = 0$;

while *not converged* **do**

 Increment t : Set $t = t + 1$;

 Set $\tilde{W}^{(0)} = W^{(t-1)}$;

for $i = 1, \dots, n$ **do**

 Assignment: Compute winning neuron $\ell = h^{\text{miss}}(x_i^{\text{obs}}, m_i, \tilde{W}^{(i-1)})$;

 Update code vectors:

for $k = 1, \dots, K$ **do**

$$\quad \quad \quad \left| \quad w_k^{(i)} = \tilde{w}_k^{(i-1)} + \varepsilon_t V_{\lambda_t}(k, \ell) \left((x_i^{\text{obs}}, x_i^{*(t-1)}) - \tilde{w}_k^{(i-1)} \right) . \right.$$

end

end

 Set $W^{(t)} = \tilde{W}^{(n)}$;

 Update imputed values: for i, j such that $m_{i,j} = 0$,

$$x_{i,j}^{*(t)} = \frac{\sum_{k=1}^K V_{\lambda_t}(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W^{(t)})) w_{k,j}^{(t)}}{\sum_{k=1}^K V_{\lambda_t}(k, h^{\text{miss}}(x_i^{\text{obs}}, m_i, W^{(t)}))}.$$

end

Output: Code vectors $W^{(t)}$ and imputed data $X^{*(t)}$.

	5%			20%			40%		
	MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
Topographic error									
basic missSOM	0.337 (0.047)	0.345 (0.035)	0.336 (0.037)	0.313 (0.032)	0.355 (0.063)	0.339 (0.046)	0.294 (0.058)	0.303 (0.057)	0.306 (0.083)
accelerated missSOM	0.352 (0.039)	0.345 (0.049)	0.338 (0.030)	0.301 (0.035)	0.299 (0.039)	0.322 (0.042)	0.282 (0.035)	0.280 (0.059)	0.285 (0.040)
Quantization error									
basic missSOM	0.405 (0.115)	0.394 (0.042)	0.407 (0.145)	0.327 (0.115)	0.326 (0.116)	0.330 (0.112)	0.225 (0.074)	0.233 (0.080)	0.233 (0.077)
accelerated missSOM	0.413 (0.149)	0.408 (0.139)	0.407 (0.138)	0.351 (0.117)	0.340 (0.116)	0.341 (0.115)	0.261 (0.081)	0.245 (0.071)	0.260 (0.074)
Imputation error									
basic missSOM	0.601 (0.140)	0.599 (0.154)	0.613 (0.171)	0.659 (0.125)	0.686 (0.141)	0.759 (0.108)	0.758 (0.093)	0.795 (0.095)	0.871 (0.091)
accelerated missSOM	0.615 (0.097)	0.624 (0.096)	0.654 (0.103)	0.665 (0.092)	0.694 (0.087)	0.750 (0.075)	0.742 (0.069)	0.793 (0.069)	0.848 (0.072)
ARI									
basic missSOM	0.922 (0.092)	0.929 (0.089)	0.937 (0.071)	0.849 (0.088)	0.846 (0.108)	0.823 (0.089)	0.686 (0.108)	0.640 (0.101)	0.623 (0.086)
accelerated missSOM	0.923 (0.090)	0.923 (0.086)	0.922 (0.082)	0.841 (0.092)	0.845 (0.081)	0.805 (0.090)	0.647 (0.094)	0.629 (0.109)	0.591 (0.089)
Computing time									
basic missSOM	33.169 (0.546)	32.940 (0.737)	31.860 (0.396)	33.076 (0.517)	32.988 (0.776)	32.948 (0.530)	33.031 (0.583)	31.920 (0.061)	32.846 (0.0644)
accelerated missSOM	0.394 (0.010)	0.387 (0.006)	0.376 (0.009)	0.557 (0.008)	0.561 (0.017)	0.556 (0.006)	0.790 (0.016)	0.753 (0.011)	0.784 (0.022)

Table 1: Comparison of the basic missSOM Algorithm 2 and its accelerated version Algorithm 3 in terms of different errors and computing time (in seconds) on the gaussian mixture data.