A Nonlinear Approach for the Determination of Missing Values in Temporal Databases

Antti Sorjamaa[1]**, Paul Merlin**[2]**, Bertrand Maillet**[2] **and Amaury Lendasse**[1]

1- Helsinki University of Technology
Laboratory of Computer and Information Science
P.O. Box 5400, 02015 HUT - Finland

2- A.A.Advisors-QCG (ABN AMRO),
Variances and Paris-1 University $PSE/CNRS$
106 bv de l'hôpital F-75647 Paris cedex 13 - France

RÉSUMÉ. *L'existence de valeurs manquantes dans les séries temporelles est un problème récurrent lors de l'utilisation de modèles financiers. En effet, de tels modèles requièrent que les bases soient cylindrées et complètes. De plus, de nombreuses bases de données financières contiennent des valeurs manquantes. Ce papier présente une nouvelle technique pour le recouvrement des valeurs manquantes. Cette méthode utilise deux techniques de projection : une non-linéaire (Cartes de Kohonen) et une linéaire (Fonction Orthogonale Empirique). La méthodologie globale présentée combine les avantages des deux méthodes pour obtenir des candidats aux valeurs manquantes. La méthode est appliquée à deux bases de données financières.*

ABSTRACT. *The presence of missing data in the underlying time series is a recurrent problem for market models. Such models make it necessary to deal with cylindrical and complete samples. Moreover, many financial databases contain missing values. This paper presents a new method for the missing values recovery. The new method is based on two projection methods: a nonlinear one (Self-Organizing Maps) and a linear one (Empirical Orthogonal Functions). The presented global methodology combines the advantages of both methods to get accurate approximations for the missing values. The methods are applied to three financial datasets.*

MOTS-CLÉS : *Valeurs manquantes, Cartes de Kohonen, SOM, Fonctions orthogonales empiriques, EOF.*

KEYWORDS: *Missing values, Self-Organizing Maps, SOM, Empirical Orthogonal Functions, EOF.*

^{e} soumission à *EJESS 2009*, le 23 mars 2009.

1. Introduction

Academics as well as practitioners often face the problem of missing data in financial timeseries. Non-quotation date, too recent inception date, intention not to report a bad performance or mistake of data provider are some of the reasons why missing values occur recurrently in financial databases. Moreover, in order to achieve good performance, most financial models need complete and cylindrical samples. Thus, most of the time, imputation methods have to be applied before running the model.

A number of methods have been developed to solve the problem and fill the missing values, both commercial and academical. The methods in both sectors can be classified into two distinct categories : deterministic methods and stochastic methods.

Self-Organizing Maps [KOH 95] (SOM) aim to ideally group homogeneous individuals, highlighting a neighborhood structure between classes in a chosen lattice. The SOM algorithm is based on unsupervised learning principle where the training is entirely stochastic, data-driven. No information about the input data is required. Recent approaches propose to take advantage of the homogeneity of the underlying classes for data completion purposes [WAN 03]. Furthermore, the SOM algorithm allows projection of high-dimensional data to a low-dimensional grid. Through this projection and focusing on its property of topology preservation, SOM allows nonlinear interpolation for missing values.

Empirical Orthogonal Functions (EOF) [PRE 88] are deterministic models, enabling linear projection to high-dimensional space. They have also been used to develop models for finding missing data [BOY 94]. Moreover, EOF models allow continuous interpolation of missing values, but are sensitive to the initialization.

This paper describes a new method, which combines the advantages of both the SOM and the EOF. The nonlinearity property of the SOM is used as a denoising tool and then continuity property of the EOF method is used to recover missing data efficiently.

The SOM is presented in Section 2, the EOF in Section 3 and the global methodology SOM+EOF in Section 4. Section 5 presents the experimental results using three financial datasets.

2. Self-Organizing Map

The SOM algorithm is based on an unsupervised learning principle, where training is entirely data-driven and no information about the input data is required [KOH 95]. Here we use a 2-dimensional network, compound in c units (or code vectors) shaped as a square *lattice*. Each unit of a network has as many weights as the length T of the learning data samples, $\mathbf{x}_n,\ n = 1, 2, ..., N$. All units of a network can be collected to a weight matrix $\mathbf{m}(t) = [\mathbf{m}_1(t), \mathbf{m}_2(t), ..., \mathbf{m}_c(t)]$ where $\mathbf{m}_i(t)$ is the T-dimensional weight vector of the unit i at time t and t represents the steps of the learning process. Each unit is connected to its neighboring units through neighborhood function $\lambda(\mathbf{m}_i, \mathbf{m}_j, t)$, which defines the shape and the size of the neighborhood at time t. Neighborhood can be constant through the entire learning process or it can change in the course of learning.

Learning starts by initializing the network node weights randomly. Then, for randomly selected sample $\mathbf{x}_{t+1}$, we calculate a Best Matching Unit (BMU), which is the neuron whose weights are closest to the sample. BMU calculation is defined as

$$\mathbf{m}_{BMU(\mathbf{x}_{t+1})} = \arg \min_{\mathbf{m}_i, i \in I} \left\{ \|\mathbf{x_{t+1}} - \mathbf{m}_i\,(t)\| \right\}, \tag{1}$$

where $I = [1, 2, ..., c]$ is the set of network node indices, BMU denotes the index of the best matching node and $\|.\|$ is standard Euclidean norm.

If the randomly selected sample includes missing values, the BMU cannot be solved outright. Instead, an adapted SOM algorithm, proposed by Cottrell and Letrémy [COT 05], is used. The randomly drawn sample $\mathbf{x_{t+1}}$ having missing value(s) is split into two subsets $\mathbf{x}_{t+1}^T = NM_{\mathbf{x_{t+1}}} \cup M_{\mathbf{x_{t+1}}}$, where $NM_{\mathbf{x_{t+1}}}$ is the subset where the values of $\mathbf{x_{t+1}}$ are not missing and $M_{\mathbf{x_{t+1}}}$ is the subset where the values of $\mathbf{x_{t+1}}$ are missing. We define a norm on the subset $NM_{\mathbf{x_{t+1}}}$ as

$$\|\mathbf{x_{t+1}} - \mathbf{m}_i\,(t)\|_{NM_{\mathbf{x_{t+1}}}} = \sum_{k \in NM_{\mathbf{x_{t+1}}}} \left(\mathbf{x}_{t+1,k} - \mathbf{m}_{i,k}(t)\right)^2, \tag{2}$$

where $\mathbf{x}_{t+1,k}$ for $k = [1, ..., T]$ denotes the k^{th} value of the chosen vector and $\mathbf{m}_{i,k}(t)$ for $k = [1, ..., T]$ and for $i = [1, ..., c]$ is the k^{th} value of the i^{th} code vector.

Then the BMU is calculated with

$$\mathbf{m}_{BMU(\mathbf{x}_{t+1})} = \arg \min_{\mathbf{m}_i, i \in I} \left\{ \|\mathbf{x_{t+1}} - \mathbf{m}_i\,(t)\|_{NM_{\mathbf{x_{t+1}}}} \right\}. \tag{3}$$

When the BMU is found the network weights are updated as

$$\mathbf{m}_i\,(t+1) = \mathbf{m}_i\,(t) - \varepsilon(t)\lambda\left(\mathbf{m}_{BMU(\mathbf{x}_{t+1})}, \mathbf{m}_i, t\right)\left[\mathbf{m}_i\,(t) - \mathbf{x_{t+1}}\right], \forall i \in I, \tag{4}$$

where $\varepsilon(t)$ is the adaptation gain parameter, which is $]0, 1[$-valued, decreasing gradually with time. The number of neurons taken into account during the weight update depends on the neighborhood function $\lambda(\mathbf{m}_i, \mathbf{m}_j, t)$. The number of neurons, which need the weight update, usually decreases with time.

After the weight update the next sample is randomly drawn from the data matrix and the procedure started again by finding the BMU of the sample. The recursive learning procedure is stopped when the SOM algorithm has converged.

Once the SOM algorithm has converged, we obtain some clusters containing our data. Cottrell and Letrémy proposed to fill the missing values of the dataset by the coordinates of the code vectors of each BMU as natural first candidates for missing value completion :

$$\pi_{(M_{\mathbf{x}})}\left(\mathbf{x}\right) = \pi_{(M_{\mathbf{x}})}\left(\mathbf{m}_{BMU(\mathbf{x})}\right), \tag{5}$$

where $\pi_{(M_{\mathbf{x}})}\left(.\right)$ replaces the missing values $M_{\mathbf{x}}$ of sample $\mathbf{x}$ with the corresponding values of the BMU of the sample. The replacement is done for every data sample and then the SOM has finished filling the missing values in the data.

The procedure is summarized in Table 1. There is a toolbox available for performing the SOM algorithm in [URL 01].

Tableau 1. *Summary of the SOM algorithm for finding the missing values.*

1 SOM node weights are initialized randomly

2 SOM learning process begins

3 Input $\mathbf{x}$ is drawn from the learning data set $\mathbf{X}$
- 3.1 If $\mathbf{x}$ does not contain missing values, BMU is found according to Equation 1
- 3.2 If $\mathbf{x}$ contains missing values, BMU is found according to Equation 3

4- Once the learning process is done, for each observation containing missing values, the weights of the BMU of the observation are substituted for missing values

3. Empirical Orthogonal Functions

This section presents Empirical Orthogonal Functions (EOF) [PRE 88]. In this paper, EOF are used as a denoising tool and for finding the missing values at the same time [BOY 94].

The EOF are calculated using standard and well-known Singular Value Decomposition (SVD)

$$\mathbf{X} = \mathbf{UDV}^{*} = \sum_{k=1}^{K} \rho_k \mathbf{u}_k \mathbf{v}_k, \tag{6}$$

where $\mathbf{X}$ is 2-dimensional data matrix, $\mathbf{U}$ and $\mathbf{V}$ are collections of singular vectors $\mathbf{u}$ and $\mathbf{v}$ in each dimension respectively, $\mathbf{D}$ is a diagonal matrix with the singular values ρ in its diagonal and K is the smaller dimension of $\mathbf{X}$ (or the number of nonzero singular values if $\mathbf{X}$ is not full rank). The singular values and the respective vectors are sorted to decreasing order.

When EOF are used to denoise the data, not all singular values and vectors are used to reconstruct the data matrix. Instead, it is assumed that the vectors corresponding to larger singular values contain more data with respect to the noise than the ones corresponding to smaller values [PRE 88]. Therefore, it is logical to select q largest singular values and the corresponding vectors and reconstruct the denoised data matrix using only them.

In the case where $q < K$, the reconstructed data matrix is obviously not the same than the original one. The larger q is selected, the more original data, which also includes more noise, is preserved. The optimal q is selected using validation methods, for example [LEN 03].

EOF (or SVD) cannot be directly used with databases including missing values. The missing values must be replaced by some initial values in order to use the EOF. This replacement can be for example the mean value of the whole data matrix $\mathbf{X}$ or the mean in one direction, row wise or column wise. The latter approach is more logical when the data matrix has some temporal or spatial structure in its columns or rows.

After the initial value replacement the EOF process begins by performing the SVD and the selected q singular values and vectors are used to build the reconstruction. In order not to lose **any** information, only the missing values of $\mathbf{X}$ are replaced with the values from the reconstruction. After the replacement, the new data matrix is again broken down to singular values and vectors with the SVD and reconstructed again. The procedure is repeated until convergence criterion is fulfilled.

The procedure is summarized in Table 2.

Tableau 2. *Summary of the EOF method for finding missing values.*

```
1    Initial values are substituted into missing values of the original data matrix $\mathbf{X}$

2    For each $q$ from 1 to $K$
     2.1   SVD algorithm calculates $q$ singular values and eigenvectors
     2.2   A number of values and vectors are used to make the reconstruction
     2.3   The missing values from the original data are filled with the values from
           the reconstruction
     2.4   If the convergence criterion is fulfilled, the validation error is calculated
           and saved and the next $q$ value is taken under inspection. If not,
           then we continue from step 2.1 with the same $q$ value

3    The $q$ with the smallest validation error is selected and used to reconstruct
     the final filling of the missing values in $\mathbf{X}$
```

4. Global Methodology

The two methodologies presented in the previous two sections are combined and the global methodology is presented. The SOM algorithm for missing values is first ran through performing a nonlinear projection for finding the missing values. Then, the result of the SOM estimation is used as initialization for the EOF method. The global methodology is summarized in Table 1.

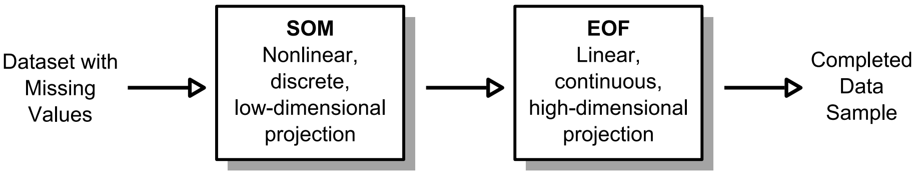

Figure 1. *Global methodology, the SOM+EOF, summarized.*

For the SOM we must select the optimal grid size c and for the EOF the optimal number of singular values and vectors q to be used. This is done using validation, using the same validation set for all combinations of the parameters c and q. Finally, the combination of SOM and EOF that gives the smallest validation error is used to perform the final filling of the data.

Even the SOM as well as the EOF are able to fill the missing values alone, the experimental results demonstrate that together the accuracy is better. The fact that these two algorithms suit well together is not surprising. Two approaches can be considered to understand the complementarity of the algorithms.

Firstly, the SOM algorithm allows nonlinear projection. In this sense, even for dataset with complex and nonlinear structure, the SOM code vectors will succeed to capture the nonlinear characteristics of the inputs. However, the projection is done on a low-dimensional grid (in our case two-dimensional) with the possibility of losing the intrinsic information of the data.

The EOF method is based on a linear transformation using the Singular Value Decomposition. Because of the linearity of the EOF approach, it will fail to reflect the nonlinear structures of the dataset, but the projection space can be as high as the dimension of the input data and remain continuous.

There is a toolbox for performing the SOM+EOF in [URL 02].

5. Experimental Results

To illustrate the accuracy of the presented methodology, we run several experiments on three financial return databases. The first one recovers the missing values when they are missing at random, the second experiment has missing values only at the beginnings of several timeseries and the third one is a publicly available financial dataset used with random missing values.

In comparison, we also experiment with a widely used methodology called Expectation Conditional Maximization, which is briefly presented in the following.

5.1. *Expected Maximization Methods*

As a benchmark to estimate how well our combination methodology performs, we choose to compare our results with those obtained by the Expectation Maximization (EM) algorithm.

The EM algorithm presented by Dempster, Laird and Rubin in 1977, is a technique to find maximum likelihood estimates in the missing data situation. Since the estimates of the mean and the covariance matrix of an incomplete dataset depend on the unknown missing values, and, conversely, estimates of the missing values depend on the unknown statistics of the data, this estimation problem is non-linear and has to be done iteratively.

The EM algorithm consists of two steps :

1) E-step calculates the expectation of the complete data sufficient statistics given the observed data and current parameter estimates.

2) M-step updates the parameter estimates through the maximum likelihood approach based on the current values of the complete sufficient statistics.

The algorithm proceeds in an iterative manner until the difference between the last two consecutive parameter estimates converges to a specified criterion. The final E-step calculates the expectation of each missing value given the final parameter estimates and the observed data. This result will be used as the imputation value.

For each iteration (t), the E-step consist of

$$Q\left(\theta\left|\theta^{(t)}\right.\right) = E\left[L\left(\theta\left|Y\right.\right)\left|Y_{obs}, \theta^{(t)}\right.\right], \tag{7}$$

where

$$\begin{cases} L\left(.\left|Y\right.\right) \text{ denotes the likelihood function conditionally to the sample,} \\ \theta \text{ the vector of parameter to be estimated,} \\ Y_{obs} \text{ the non missing values,} \\ Y \text{ the sample,} \\ \theta^{(t)} \text{ the last vector of parameter estimated.} \end{cases}$$

Then the $(t+1)^{th}$ M-step finds $\theta^{(t+1)}$ to maximize $Q\left(\theta\left|\theta^{(t)}\right.\right)$ such that

$$Q\left(\theta^{(t+1)}\middle|\theta^{(t)}\right) = \max_{\theta} Q\left(\theta\middle|\theta^{(t)}\right). \tag{8}$$

The main drawback of the EM algorithm is when the M-step is not in closed form. In this case, the M-step could be difficult to perform.

Meng and Rubin [1993] proposed an alternative algorithm called the Expectation Conditional Maximization (ECM) to solve this problem. The M-step is decomposed in multiple conditional maximization. Consider $\theta = [\theta_1, \theta_2, ..., \theta_k]$ a k-dimensional vector of parameters. Then the CM-step consist in k successive maximizations, with previous notation

$$Q\left(\theta^{(t+1)}\middle|\theta^{(t)}\right) = \max_{\theta_i} Q\left(\theta\middle|\theta^{(t)}\right), \text{ for } i = 1, ..., k. \tag{9}$$

Otherwise, the ECM algorithm performs in the same way than the EM algorithm presented before.

5.2. *North American Fund Returns*

For the first experiment, we use a dataset of North American fund returns[1] composed with 679 funds on a 4-year period of 219 weekly values, which give a total of 148 701 values. Then, in the definition of the dataset $\mathbf{X}$, the size of the dimensions is $T \times N$ which is equal to 219×679.

The fund return correspond to the yield of asset values between two consecutive dates as

$$r_t = \frac{v_{t+1}}{v_t} - 1, \tag{10}$$

where v_t is the value of the considered asset at time t.

There are no missing values contained in the original database. Figure 2 shows 10 rescaled fund values $\left(v_t^{'} = 100 \prod_{i=1}^{t} (1 + r_t)\right)$. The fund values are correlated time series including first order trends.

Before running any experiments, we randomly remove for testing purposes 7.5 percent of the data, which corresponds to 11 152 missing values. For each validation set, the same amount of data is removed from the dataset. Therefore, for the model selection and learning we have a database with a total of 15 percent of missing values.

We use Monte Carlo Cross-Validation method with 10 folds to select the optimal parameters for the SOM, the EOF and the SOM+EOF. The 10 selected validation sets are the same for each method and the validation results are presented in the following.

[1] Data provided by Lipper, A Reuters Company.

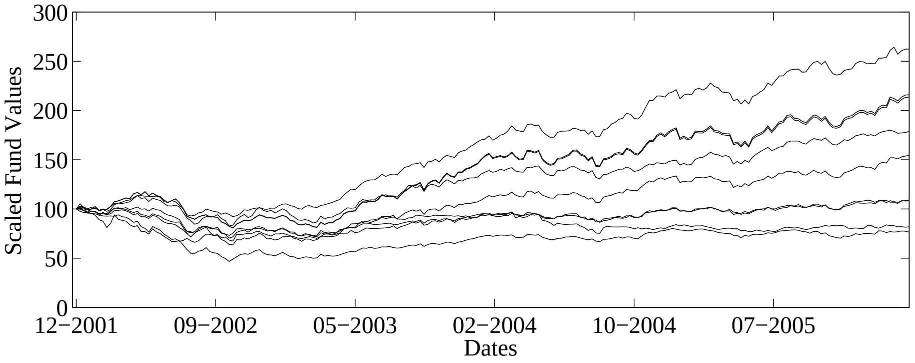

Figure 2. *Rescaled asset values of 10 funds present in the database. Source : Lipper ; North American Fund Weekly Return from 28/12/2001 to 03/03/2006. Computation from the Authors.*

5.2.1. *SOM*

Focusing on the topology preservation property of the SOM algorithm, we project our data on a large sized map. For each grid size, we compute the Root Mean Square Errors (RMSE) of the reconstruction on all validation sets. Then the grid size giving the smallest validation error is selected and the corresponding grid size is used to make the final filling. The validation errors are shown in Figure 3.

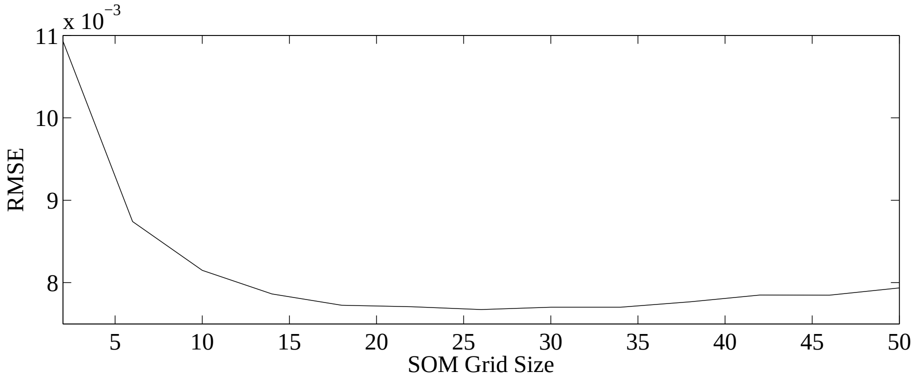

Figure 3. *Validation errors with respect to square number of grid size using the SOM method. Source : Lipper ; North American Fund Weekly Return from 28/12/2001 to 03/03/2006. Computation from the Authors.*

The optimal size of the SOM grid is found to be 26×26, which is a total of 676 units, see Figure 3. Therefore, we have more code vectors in the SOM than observations (629). It means that we have a nonlinear interpolation between the observations and better approximation of the missing values.

Once the optimal grid size is found, we apply the SOM algorithm and fill in all the missing values. Now we have only 7.5 percent of the data missing due to the removed test set. The test and validation errors are summarized in the end of the section, in Table 3.

5.2.2. *EOF*

The validation errors with respect to q for the EOF method are shown in Figures 4 and 5. In this case, when the EOF is used alone, the missing values are initialized using the column mean of the dataset calculated with only known values of each column.

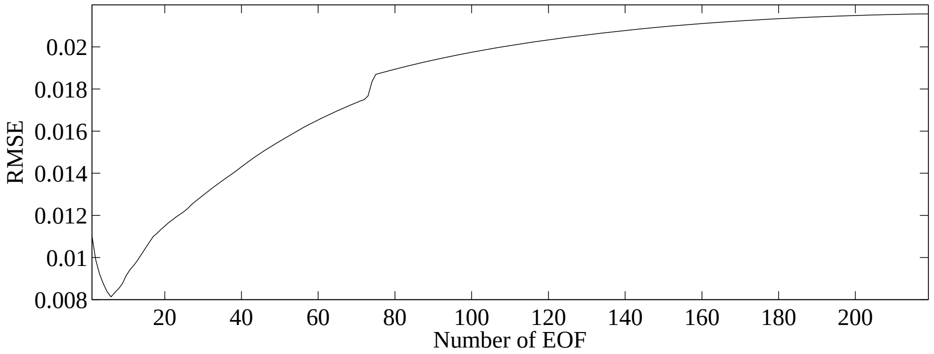

Figure 4. *Validation errors with respect to the number of EOF with the plain column mean initialization. Source : Lipper ; North American Fund Weekly Return from 28/12/2001 to 03/03/2006. Computation from the Authors.*

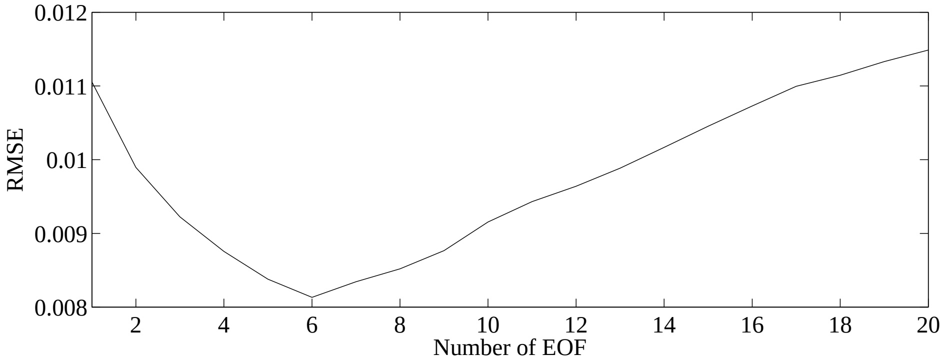

Figure 5. *EOF validation errors zoomed. Source : Lipper ; North American Fund Weekly Return from 28/12/2001 to 03/03/2006. Computation from the Authors.*

From the Figure 5 the smallest error is achieved with q equal to 6. This number of EOF is relatively small compared to the maximum of 219 EOF. It suggests quite strong noise influence in the data and that there is only a small number of efficient EOF needed to represent the denoised data.

5.2.3. *SOM+EOF*

In our experiments, we have seen that it is not enough to select the SOM grid size and the number of EOF separately. Instead, both parameters must be optimized together, simultaneously. Even though this increases the computational load, it gives more accurate results.

In Figures 6 and 7 the validation RMSEs are presented. The first figure shows the minimum EOF errors with respect to the SOM grid size and the latter figure the EOF errors with the selected SOM grid size.

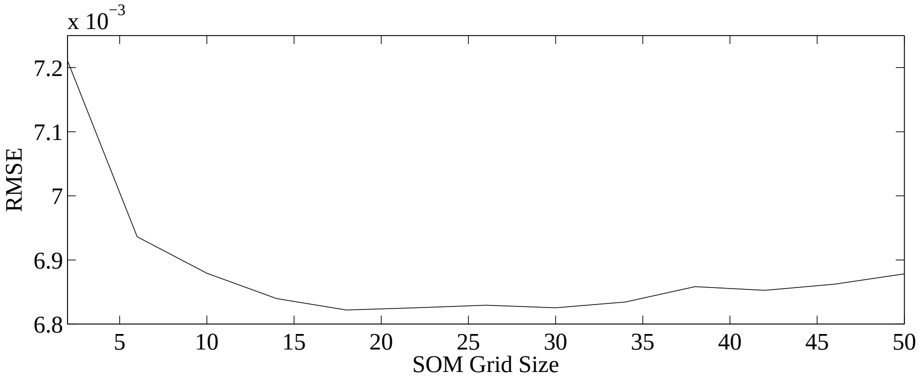

Figure 6. *Validation errors with respect to the SOM grid size using the SOM+EOF.*
Source : Lipper ; North American Fund Weekly Return from 28/12/2001 to 03/03/2006. Computation from the Authors.

From the Figures 6 and 7 the smallest error is achieved with the SOM grid size equal to 18×18 and the number of EOF q equal to 40.

The number of selected EOF is larger with SOM initialization than with the column mean initialization. It suggests there are more efficient EOF to use in the approximation of the missing values than with the plain column mean initialization and that the SOM has already denoised the data.

The SOM size is decreased when compared to the SOM method alone. It suggests that the nonlinear interpolation is not as crucial than using the SOM alone, but instead the denoising property is enhanced by limiting the number of SOM nodes.

It is also evident that the individual optimization of the parameters is not guaranteeing appropriate performance, which can be seen from totally different selections of parameters when using the SOM+EOF than the methods individually.

Table 3 summarizes the errors of the SOM, the EOF and the the SOM+EOF methods.

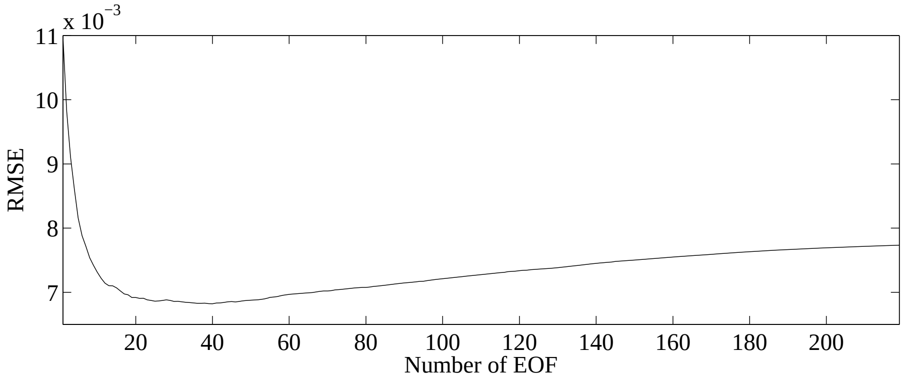

Figure 7. *Validation errors with respect to the number of EOF with the SOM grid size 18×18. Source : Lipper ; North American Fund Weekly Return from 28/12/2001 to 03/03/2006. Computation from the Authors.*

Tableau 3. *Validation and test RMS errors for all the methods. Source : Lipper ; North American Fund Weekly Return from 28/12/2001 to 03/03/2006. Computation from the Authors.*

10^{-3}	Validation Error	Test Error
ECM	13.8	13.6
SOM	7.67	7.33
EOF	8.13	7.83
SOM+EOF	6.82	6.59

From the Table 3, we can see that the SOM+EOF outperforms the EOF reducing the validation and test errors by 16 percent and the SOM errors more than 10 percent.

5.2.4. *More Missing Values*

In order to test the robustness of the SOM+EOF method, we experiment the effect of increasing the percentage of missing values in the database.

Before selecting the test or the validation sets, we randomly remove 30 percent of the data. Then the same procedure as before is performed by first removing 7.5 percent of the remaining data for the test set and then for each validation set another 7.5 percent.

Finally, the total amount of missing data in the learning phase is around 42 percent, which makes the missing value problem considerably harder than in the previous experiments.

The validation RMS errors for the SOM method are shown in Figure 8.

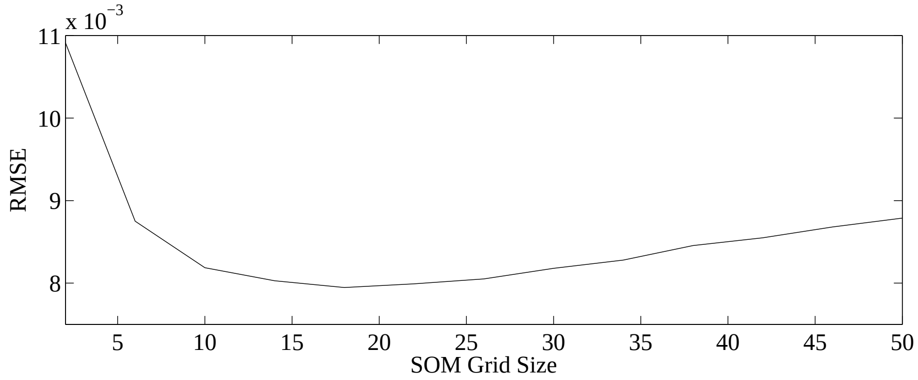

Figure 8. *Validation errors with respect to square number of grid size using the SOM method. Source : Lipper ; North American Fund Weekly Return from 28/12/2001 to 03/03/2006. Computation from the Authors.*

From Figure 8 the SOM grid size with the smallest RMS error is 18×18, which is smaller than previously using the SOM method. It means that when the percentage of missing values increases, the need for the SOM nodes decrease as there is less data to use in the interpolation of the missing values.

The validation errors for the SOM+EOF method are presented in Figures 9 and 10.

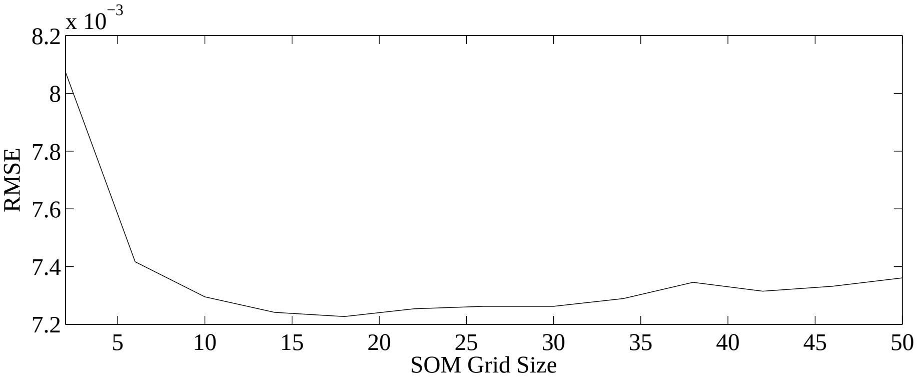

Figure 9. *Validation errors with respect to the SOM grid size using the SOM+EOF. Source : Lipper ; North American Fund Weekly Return from 28/12/2001 to 03/03/2006. Computation from the Authors.*

From Figure 9 the optimal SOM size is selected to 18×18, which is the same size than using the SOM alone. It means, that the SOM method alone is definitely not accurate enough to perform the filling of missing values alone. Therefore, it is not possible to enhance the noise removal power over interpolation performance in this case.

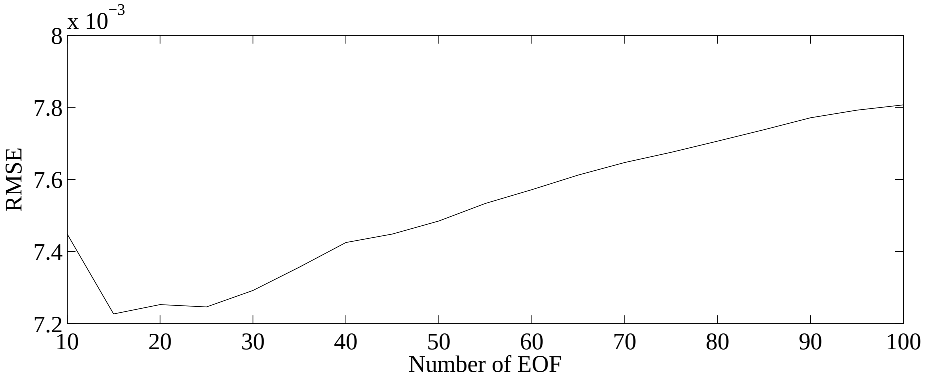

Figure 10. *Validation errors with respect to the number of EOF with the SOM grid size 18×18. Source : Lipper ; North American Fund Weekly Return from 28/12/2001 to 03/03/2006. Computation from the Authors.*

From Figure 10 the optimal number of EOF is found to be 15, which is less than in the case with less missing values. The smaller number of EOF is explained by the fact that the increased number of missing values creates more uncertainty and, therefore, the smaller singular values and the related vectors become more and more unusable in the reconstruction process.

The validation and test errors are summarized in Table 4.

Tableau 4. *Validation and test RMS errors for all the methods. Source : Lipper ; North American Fund Weekly Return from 28/12/2001 to 03/03/2006. Computation from the Authors.*

10^{-3}	Validation Error	Test Error
ECM	9.50	13.8
SOM	7.94	7.73
EOF	9.07	9.17
SOM+EOF	7.22	7.01

From Table 4 it can be seen that the SOM+EOF method has decreased the validation and test errors both by 9 percent compared to the SOM method. The improvement is slightly worse than in the case of less missing values. Still, there is notable performance upgrade when using the SOM+EOF method.

Comparing the error values above with the values in Table 3, we can see that all errors are increased roughly the same amount, except the ECM method. The validation error is significantly better with more missing values than with less. However, the test error is higher in comparison with the previous case as well as with the SOM, EOF and SOM+EOF methodologies.

Based on the findings, it can be concluded that the SOM and EOF based filling methods are robust and can handle efficiently even large amount of missing values contained in the database.

5.3. *European Fund Returns*

For the next experiment, we focus on a different example of missing values. Rebuilding past performance of funds is a recurrent problem for financial professionals (too short funds history). Thus, we choose to rebuild the beginnings of several time series. We use a dataset of European Fund Weekly Returns[2] from 07/11/2003 to 27/10/2006 composed of 300 funds with 175 weekly values, which give a total of 52 500 values.

We randomly remove for testing purposes 10 percent of the data at the beginning of several time series. The beginning is defined as the first third of the length of the series. We constraint the random deletion process to leave at least one fourth of the time series without any missing values. For validation, 10 percent more is removed at the beginning of the remaining time series.

We apply the same validation procedures as in the previous experiments to select the optimal SOM grid size and the number of EOF.

The optimal size of the SOM grid is found to be in mean (12×12) that is 144. Once the optimal grid size is found, we apply the SOM algorithm and fill in all the missing values. When the EOF is performed, initial missing values are substituted as the column means of the original matrix. At last, SOM estimations are then used as initialization for the EOF algorithm.

The validation errors with respect to q for the EOF alone and the EOF performed with the SOM initialization are shown in Figure 4.

From the Figure 11, we note that the smallest error is achieved with q equal to 13 using the EOF with plain column mean initialization and 23 when the EOF is initialized using the SOM. Table 5 summarizes the mean errors of the three methods.

From the Table 5 we can see that the SOM+EOF outperforms the EOF and the SOM reducing the test error by 31 percent compared to the SOM and 26 percent compared to the EOF.

5.4. *Public Financial Dataset*

The third example is performed with a publicly available dataset, which can be found from [URL 02]. It is an internal database, similar to the two previous examples, but due to the confidentiality, more details cannot be given.

[2] Data provided by Standard and Poors

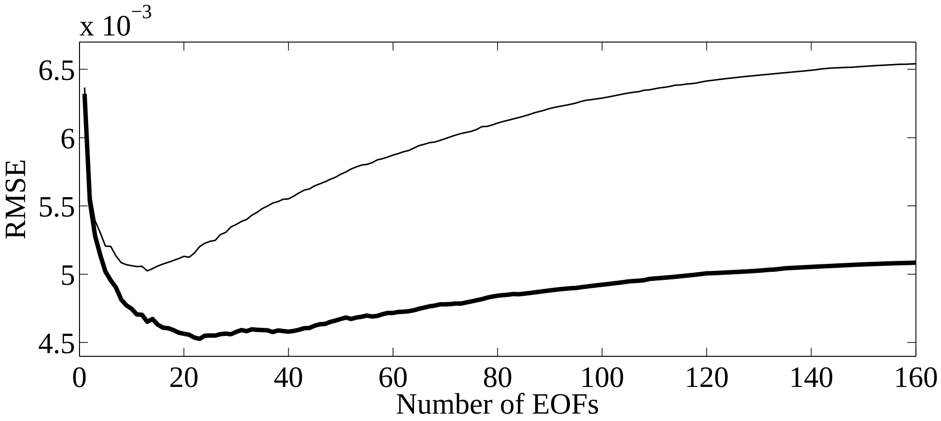

Figure 11. *Validation rrrors with respect to the number of EOF with the plain column mean initialization, the fine gray line, and with the SOM initialization, the bold black line. Source : SnP ; European Fund Weekly Return (07/11/2003 to 27/10/2006). Computation from the Authors.*

Tableau 5. *Validation and test RMS errors for all the methods. Source : SnP ; European Fund Weekly Return from 07/11/2003 to 27/10/2006. Computation from the Authors.*

10^{-3}	Validation Error	Test Error
ECM	7.98	8.25
SOM	5.09	5.06
EOF	5.02	4.85
SOM + EOF	4.53	3.83

The dataset contains 120 series, 121 values each, and it has no missing values inherently in it. The procedure is the same than before and the results are shown in Figures from 12 to 15 and the validation and test errors are summarized in Table 6.

From Figure 12 we can see that the optimal grid size for the SOM is 7×7 giving an error of 0.0417. This time the optimal number of SOM nodes 42 is clearly smaller than the number of samples 120. Also, the validation error curve is less smooth than with the other datasets, which suggests more difficult dataset to fill.

From Figure 13 we can see that the optimal number of EOF is 3 giving an error of 0.0395. The number of EOF is very small compared to the number of available EOF 120. This suggests very strong noise influence in the data and is supporting the observation of more difficult dataset to fill.

From Figure 14 the optimal SOM size is found to be 8×8. It is roughly the same than using SOM alone, but because of the EOF methodology performed after SOM initialization, the obtained error is lower, 0.0389.

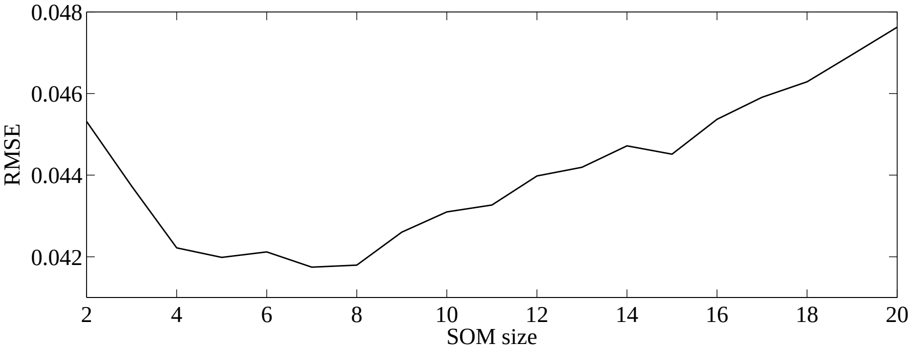

Figure 12. *Validation Errors with Respect to the size of the SOM grid.*

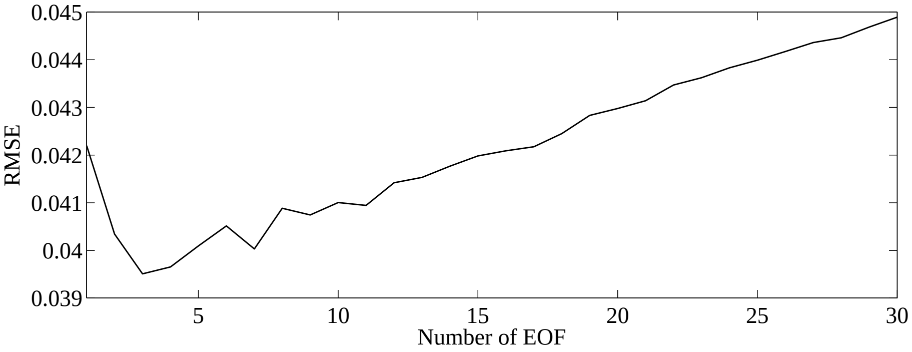

Figure 13. *Validation Errors with Respect to the number of EOF with column mean initialization.*

From Figure 15 the optimal EOF is found to be 18, which is 6 times larger when using EOF alone. This clearly demonstrates the necessity of a good initialization necessary for the EOF, and SOM can provide that.

Also, it can be noted that the number of EOF selected for the EOF alone is a clear local minimum even when using SOM initialization. It suggests that not all 18 EOF are necessary, even though the global optimum is obtained, but instead some of the EOF between 3 and 18 could be pruned out.

Comparing the validation and test errors in the Table 6, the SOM+EOF clearly outperforms the other methodologies. Since this dataset is the hardest one, all the errors are also larger by one order of magnitude compared to the American and European funds return datasets.

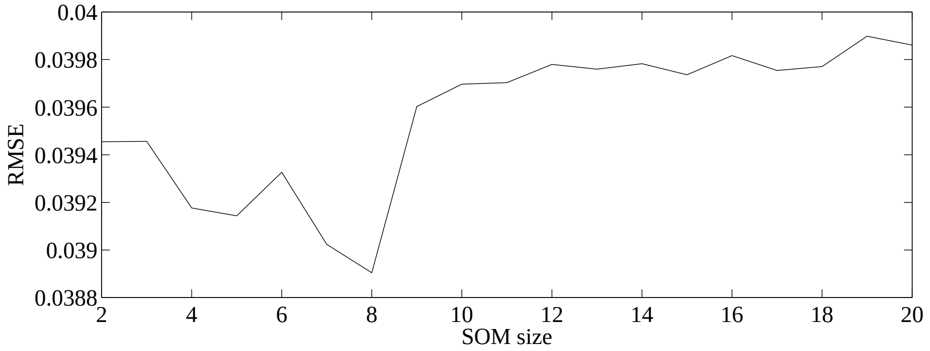

Figure 14. *Validation Errors with Respect to the size of the SOM grid.*

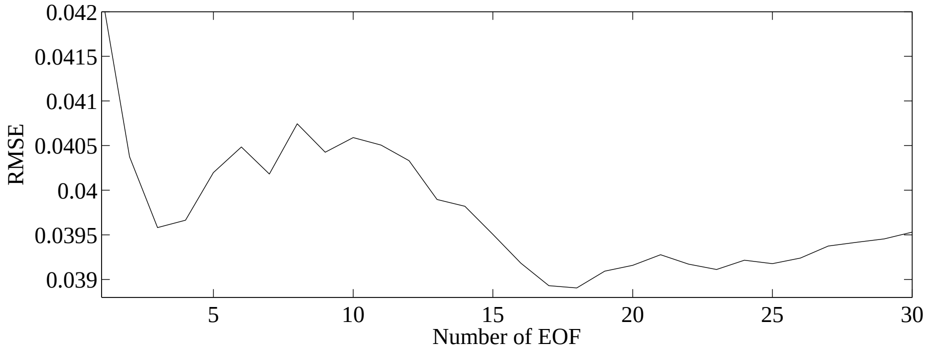

Figure 15. *Validation Errors with Respect to the number of EOF with SOM initialization using a grid size 8×8.*

Tableau 6. *Validation and test RMS errors for all the methods using a publicly available financial dataset.*

10^{-2}	Validation Error	Test Error
ECM	4.18	4.34
SOM	4.17	3.92
EOF	3.95	3.88
SOM + EOF	3.89	3.60

6. Conclusion

In this paper, we have compared 3 methods for finding missing values in temporal databases. The methods are Self-Organizing Maps (SOM), Empirical Orthogonal Function (EOF) and the combination of the two, the SOM+EOF method.

The advantages of the SOM include the ability to perform nonlinear projection of high-dimensional data to lower dimension with interpolation between discrete data points.

For the EOF, the advantages include high-dimensional linear projection of high-dimensional data and the speed and the simplicity of the method.

The combination of the two methods include the advantages of both individual methods, leading to a new accurate approximation methodology for finding the missing values. The performance obtained in test show the accuracy of the new methodology.

It has also been shown experimentally that the optimal number of code vectors used in the SOM has to be larger than the number of observations. It is necessary in order to take the advantage of the self-organizing property of the SOM and the interpolation ability for finding the missing data.

Furthermore, the amount of missing values is neither restricting the usage of the method nor seriously decreasing the performance.

For further work, the modifications and performance upgrades for the global methodology are fine-tuned for different types of datasets. The methodology will then be applied to datasets from other field of science, for example climatology and process data.

Bibliography

[BOY 94] BOYD J., KENNELLY E., PISTEK P. « Estimation of eof expansion coefficients from incomplete data. », *Deep-sea research*, vol. 41, n° 10, 1994, p. 1479-1488.

[COT 05] COTTRELL M., LETRÉMY P., « Missing values : Processing with the kohonen algorithm. », *Applied Stochastic Models and Data Analysis*, 2005, p. 17-20.

[KOH 95] KOHONEN T., *Self-Organizing Maps*, Springer-Verlag, Berlin, 1995.

[LEN 03] LENDASSE A., WERTZ V., VERLEYSEN M. « Model selection with cross-validations and bootstraps - application to time series prediction with rbfn models. », *LNCS*, vol. 2714, 2003, p. 573-580.

[PRE 88] PREISENDORFER R., *Principal Component Analysis in Meteorology and Oceanography*, Elsevier, 1988.

[URL 01] SOM TOOLBOX : HTTP ://WWW.CIS.HUT.FI/PROJECTS/SOMTOOLBOX/.

[URL 02] SOM+EOF TOOLBOX : HTTP ://WWW.CIS.HUT.FI/PROJECTS/TSP/INDEX.PHP ?PAGE=RESEARCH&SUBPAGE=DOWNLOADS.

[WAN 03] WANG S., « Application of self-organising maps for data mining with incomplete data set. », *Neural Computing and Applications*, vol. 12, n° 1, 2003, p. 42-48.