Time Series Prediction as a Problem of Missing Values: Application to ESTSP2007 and NN3 Competition Benchmarks

Antti Sorjamaa and Amaury Lendasse

***Abstract*— In this paper, time series prediction is considered as a problem of missing values. A method for the determination of the missing time series values is presented. The method is based on two projection methods: a nonlinear one (Self-Organized Maps) and a linear one (Empirical Orthogonal Functions). The presented global methodology combines the advantages of both methods to get accurate candidates for the prediction values. The methods are applied to two time series competition datasets.**

I. Introduction

The presence of missing values in the underlying time series is a recurrent problem when dealing with databases. A number of methods have been developed to solve the problem and fill the missing values. The methods can be classified into two distinct categories: deterministic methods and stochastic methods.

Self-Organizing Maps [1] (SOM) aim to ideally group homogeneous individuals, highlighting the neighborhood structure between classes in a chosen lattice. The SOM algorithm is based on an 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 the data completion purposes [2]. Furthermore, the SOM algorithm allows the projection of a high-dimensional data to a low-dimensional grid. Through this projection and focusing on its property of topology preservation, the SOM allows a nonlinear interpolation for the missing values.

Empirical Orthogonal Function (EOF) [3] models are deterministic enabling a linear projection without the loss in the data dimensionality. They have also been used to develop models for finding missing data [4]. Moreover, EOF models allow a continuous interpolation of the missing values, but are sensitive to the initialization.

This paper describes a new methodology, which combines the advantages of both the SOM and the EOF. The nonlinear interpolation property of the SOM is used as an accurate initialization tool and then the continuity property of the EOF method is used to recover missing data efficiently.

The SOM is presented in the Section III, the EOF in Section IV and the global methodology SOM+EOF in Section V. Section VI presents the experimental results using two competition datasets; The ESTSP2007 [5] and the NN3 [6] competition benchmarks.

Antti Sorjamaa and Amaury Lendasse are with Adaptive Informatics Research Centre - Helsinki University of Technology, P.O. Box 5400, 02015 HUT, Finland (Email: {Antti.Sorjamaa,Lendasse}@hut.fi)

II. Time Series Prediction

A. Data with Missing Values

In the time series prediction problem, the samples are generated by sliding a fixed window over the time series and taking each window full of values as a sample. The size of the window and thus the length of the samples is T. All samples are collected to a *regressor matrix*

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_j \end{bmatrix}, \; j = 1, 2, ..., N, \tag{1}$$

where N is the number of samples and each $\mathbf{x}_j$ is a T-dimensional sample vector.

When predicting the future of the time series, the missing values are added to the end of the known values of the time series. Then, logically the regressor matrix is missing some values in the lower right corner. The shape and the size of the area of the missing values depends on the used method and the horizon of prediction.

B. Prediction Strategy

There are three prediction strategies for the long-term prediction of time series that are mainly used. The first and the least calculation intensive is the *Recursive* prediction strategy, where the model selected in the learning phase for the first time step is used repeatedly, or recursively, as far as necessary. The predicted values are used as known values and the prediction is done always only one step at a time.

The next alternative is to use a different model to predict each time step. This *Direct* prediction strategy needs a different model for each time step and is therefore many times more calculation intensive. In many cases the Direct is still an appealing choice, because of the increased accuracy compared to the Recursive strategy. Whereas the Recursive strategy suffers from the accumulation of the prediction errors, the Direct does not.

Third alternative is to use a mix of the two, called *DirRec* prediction strategy [7]. With this prediction strategy a different model is trained for each time step and all predicted values are used as a known values in the process. It means that the regressor is increased by one in every time step, when the previous prediction is included in the learning data. This increases the calculation time in the learning process but in many cases, the accuracy is also better.

In this case, when the time series prediction is considered as a missing value problem, the whole set of values to be

predicted is estimated at once. Strictly speaking the strategy used here is none of the above, but instead an *all-at-once* strategy.

III. 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 [1]. Here we use a 2-dimensional network, composed of 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 a neighborhood function $\lambda(\mathbf{m}_i, \mathbf{m}_j, t)$, which defines the shape and the size of the neighborhood at time t. The neighborhood can be constant through the entire learning process or it can change in the course of learning.

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

$$\mathbf{m}_{BMU(\mathbf{x}_{t+1})} = \arg\min_{\mathbf{m}_i, i \in I} \{\|\mathbf{x_{t+1}} - \mathbf{m}_i(t)\|\}, \quad (2)$$

where $I = [1, 2, ..., c]$ is the set of network node indices, the BMU denotes the index of the best matching node and $\|.\|$ is a 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 [8], 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

$$\|\mathbf{x_{t+1}} - \mathbf{m}_i(t)\|_{NM_{\mathbf{x_{t+1}}}} = \sum_{k \in NM_{\mathbf{x_{t+1}}}} (\mathbf{x}_{t+1,k} - \mathbf{m}_{i,k}(t))^2, \quad (3)$$

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

$$\mathbf{m}_{BMU(\mathbf{x}_{t+1})} = \arg\min_{\mathbf{m}_i, i \in I} \left\{\|\mathbf{x_{t+1}} - \mathbf{m}_i(t)\|_{NM_{\mathbf{x_{t+1}}}}\right\}. \quad (4)$$

When the BMU is found the network weights are updated as

$$\mathbf{m}_i(t+1) = \ldots \mathbf{m}_i(t) - \varepsilon(t)\lambda\left(\mathbf{m}_{BMU(\mathbf{x}_{t+1})}, \mathbf{m}_i, t\right)[\mathbf{m}_i(t) - \mathbf{x_{t+1}}], \quad \forall i \in I, \quad (5)$$

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 is started again by finding the BMU of the sample. The 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 the missing value completion:

$$\pi_{(M_{\mathbf{x}})}(\mathbf{x}) = \pi_{(M_{\mathbf{x}})}\left(\mathbf{m}_{BMU(\mathbf{x})}\right), \quad (6)$$

where $\pi_{(M_{\mathbf{x}})}(.)$ 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 I. There is a toolbox available for performing the SOM algorithm in [9].

TABLE I
SUMMARY OF THE SOM ALGORITHM FOR FINDING THE MISSING VALUES.

1) SOM node weights are initialized randomly
2) SOM learning process begins
   a) Input $\mathbf{x}$ is drawn from the learning data set $\mathbf{X}$
      i) If $\mathbf{x}$ does not contain missing values, BMU is found according to Equation 2
      ii) If $\mathbf{x}$ contains missing values, BMU is found according to Equation 4
   b) Neuron weights are updated according to Equation 6
3) Once the learning process is done, for each observation containing missing values, the weights of the BMU of the observation are substituted for the missing values

IV. Empirical Orthogonal Functions

This section presents a method called Empirical Orthogonal Functions (EOF) [3]. In this paper, the EOF are used as a denoising tool and for finding the missing values at the same time [4].

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

$$\mathbf{X} = \mathbf{UDV}^* = \sum_{k=1}^{K} \rho_k \mathbf{u}_k \mathbf{v}_k, \quad (7)$$

where $\mathbf{X}$ is a 2-dimensional data matrix, $\mathbf{U}$ and $\mathbf{V}$ are the 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 a decreasing order.

When the 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 [3]. 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 [10].

The EOF (or the 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 a convergence criterion is fulfilled.

The procedure is summarized in Table II.

TABLE II
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
   a) SVD algorithm calculates q singular values and eigenvectors
   b) A number of values and vectors are used to make the reconstruction
   c) The missing values from the original data are filled with the values from the reconstruction
   d) 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 a) 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 $\mathbf{X}$

V. 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 Figure 1.

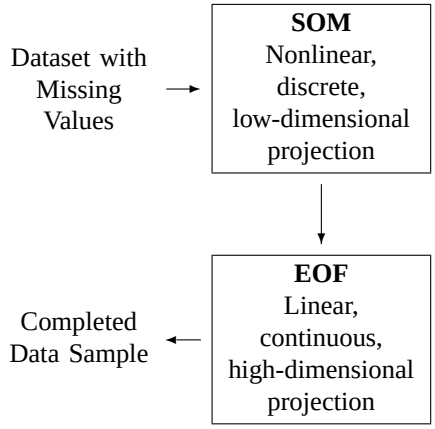

Fig. 1. Global methodology summarized.

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.

While both the SOM and 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 perspectives can be considered to understand the complementarity of the algorithms.

Firstly, the SOM algorithm allows nonlinear projection. In this sense, even for a dataset with a 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 [11].

VI. EXPERIMENTAL RESULTS

This paper presents an application of the SOM+EOF method to two time series prediction benchmarks; The ESTSP2007 competition dataset and the NN3 competition.

A. ESTSP2007 competition

This time series prediction benchmark includes a total of 875 values from an unknown origin. The dataset is shown in Figure 2. More information and the dataset can be found from the ESTSP2007 conference website [5].

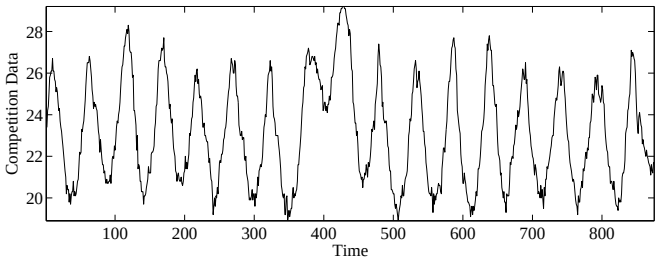

Fig. 2. ESTSP 2007 Competition dataset.

For the model selection purposes the dataset is divided into two sets, learning and validation set. The learning set consists of 465 first values and the rest belongs to the validation set. The optimal regressor size is set to 11 after many trial and error experiments.

The optimal SOM size is selected using a simple validation procedure, where the SOM learning is performed using only the learning set and the validation set is used to tune the SOM size for one step ahead prediction. The validation errors are shown in Figure 3.

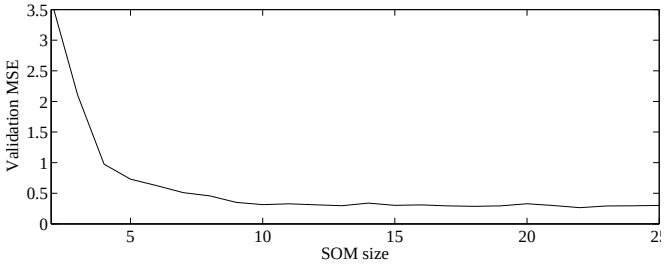

Fig. 3. Validation errors with respect to the SOM grid size.

From Figure 3 the optimal SOM size is selected to 13×13 with validation error of 0,297. There is only very small difference in the validation error with larger SOM sizes.

The only parameter of the EOF method is tuned using the same learning and validation sets than with the SOM to get comparable results. Also the regressor size is kept the same than with the SOM and the optimization is done for one step ahead prediction. The validation errors are shown in Figure 4.

From Figure 4 the optimal number of EOF is selected to 2 with validation error of 0,451. The result suggests relatively strong noise influence in the singular values after the third one, where the validation error is increasing rapidly.

For the SOM+EOF method the two separate methods are combined and the validation is performed for each

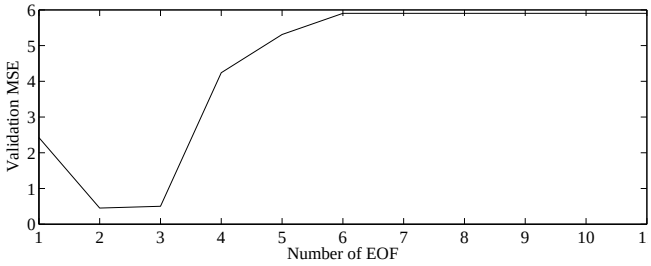

Fig. 4. Validation errors with respect to the number of EOF.

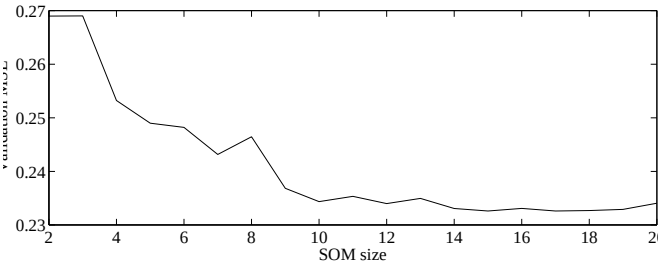

Fig. 5. Minimum validation errors with respect to the SOM size using the SOM+EOF method.

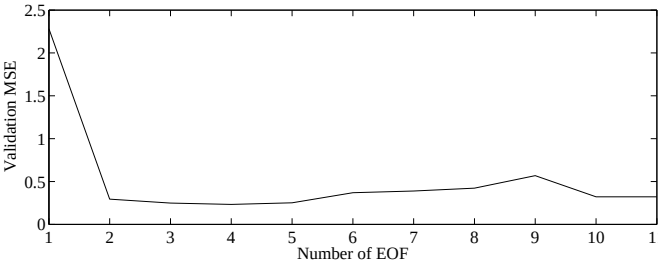

Fig. 6. Validation errors with respect to the number of EOF using SOM size 15×15.

combination of the SOM size and the number of EOF. The validation errors are shown in Figure 5 and 6.

From Figure 5 the optimal SOM is selected to be 15×15 and from Figure 6 the optimal number of EOF to 4 with the validation error of 0,233.

For one step ahead prediction the regressor size is selected to 11, but for the 50 steps ahead the regressor size is increased to 60 in order to fit the missing values to the regressor.

Our experiments with several other datasets have shown that the EOF method uses larger number of EOF when the regressor size is increased. Therefore, the final prediction is done using the number of EOF fixed to 8. The prediction of the 50 timesteps is shown in Figure 7.

From the Figure 7 it seems that that the prediction has removed the noise and is predicting the next peak of the time series quite well.

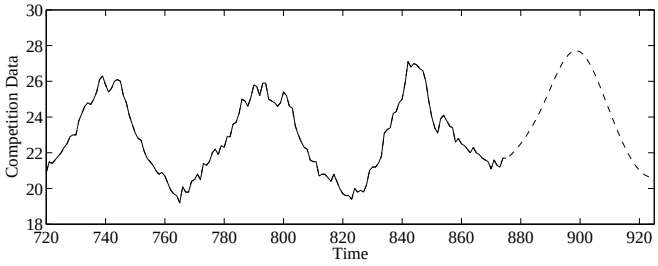

Fig. 7. Prediction of 50 next values of the competition dataset. The real values are presented by the solid line and the dashed one presents the prediction.

B. NN3 Competition

The NN3 competition consists of 11 different time series with variable lengths ranging from 126 values to 115 values. In this paper, the results with two time series are presented, namely with the 3^{rd} and the 4^{th} time series, shown in Figures 8 and 9 respectively. For more information about the competition visit [6].

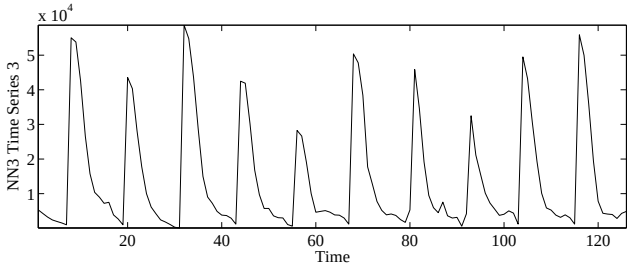

Fig. 8. NN3 Competition dataset, 3^{rd} time series.

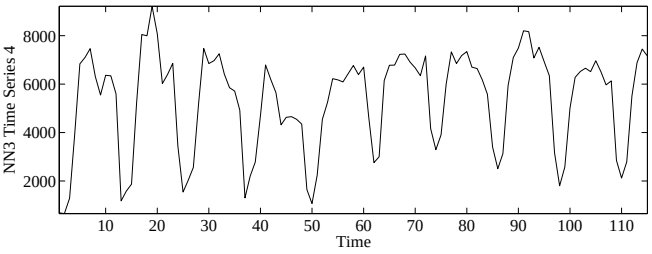

Fig. 9. NN3 Competition dataset, 4^{th} time series.

Because the EOF method was not as good as the SOM and the SOM+EOF, we use only the two latter ones with the NN3 competition time series. Also, due to the scale of the series, the normalized MSE is used in the validation error graphs. Finally, we use a 10-fold Cross-Validation instead of a simple validation in order to stabilize the parameter selection results. Otherwise, the procedure follows the one described in the previous section.

1) Time Series 3: The results for the 3^{rd} time series are presented in the following. In Figure 10 the 10-fold Cross-Validation NMSE for the SOM and the SOM+EOF method are presented. The used regressor size is 15, which is selected empirically using trial and error.

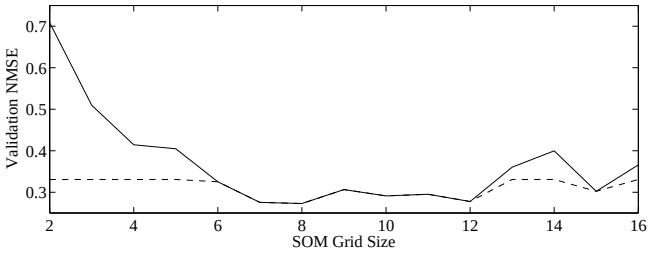

Fig. 10. Validation errors of the 3^{rd} time series. Solid line represents the SOM and the dashed one the SOM+EOF.

From Figure 10 the smallest normalized validation error is 0,27 and it is achieved with the SOM size 8×8 with the both methods. In this case, the selected number of EOF is the maximum 15. The validation NMSE is also the same than with the SOM.

Figure 11 shows the EOF validation errors using the SOM grid size 8×8.

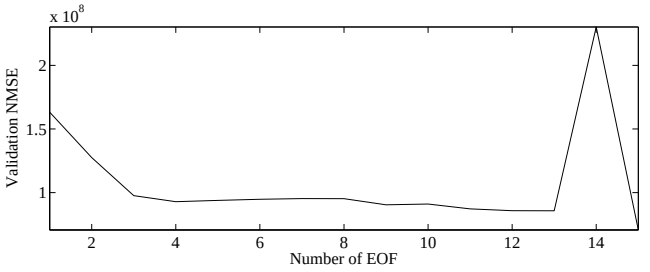

Fig. 11. EOF validation errors of the 3^{rd} time series using SOM grid size 8×8.

From Figure 11 we can clearly see, that the second last singular value contains more noise than any other value. This must be taken into account when selecting the parameters for the final prediction.

Because the regressor size must be increased to 33 from the initial 15 in order to fit the 18 missing values in the regressor, the number of EOF must also be increased. Therefore, taking into account the previous findings, the number of EOF to be used in the final prediction is fixed to 17.

The final prediction using the SOM+EOF method is shown in Figure 12.

2) Time Series 4: The results for the 4^{th} time series are presented in the following. In Figure 13 the 10-fold Cross-Validation errors with the SOM and the SOM+EOF are presented. The regressor size is set to 13 after several trial and error experiments.

From Figure 13 the SOM size with the lowest validation error is 8×8 for the SOM method and 11×11 for the

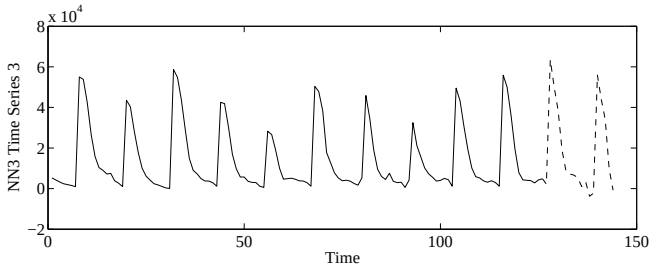

Fig. 12. Prediction of the 3^{rd} time series. Solid line represents the known time series and the dashed one the prediction using the SOM+EOF method.

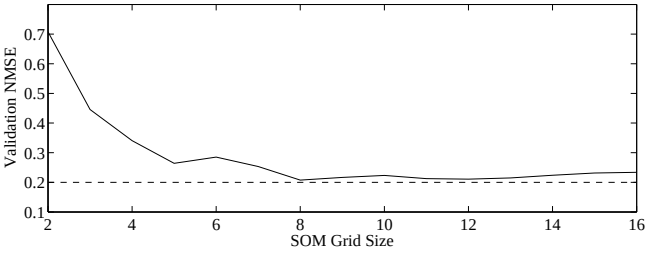

Fig. 13. Validation errors of the 4^{th} time series. Solid line represents the SOM and the dashed one the SOM+EOF.

SOM+EOF. The NMSE for the SOM is 0,21 and for the SOM+EOF 0,20. The number of EOF for the selected SOM size is 5.

For the prediction, the regressor size is increased to 31 from the initial 13 in order to fit the 18 missing values in the regressor.

Similarly than before, the number of EOF must also be increased. The final number of EOF is fixed to 8. The prediction of 18 timesteps is shown in Figure 14.

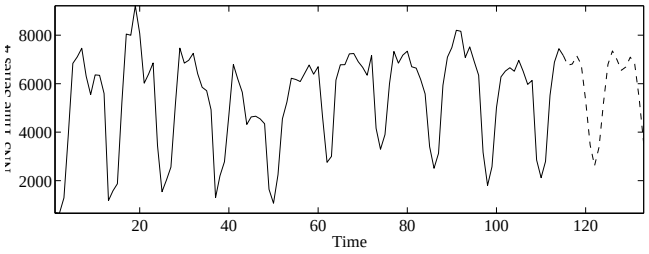

Fig. 14. Prediction of the 4^{th} time series. Solid line represents the known time series and the dashed one the prediction.

VII. Conclusion

In this paper, we have presented a new methodology for finding the missing values in a temporal database. The methodology combines the Self-Organizing Maps (SOM) and the Empirical Orthogonal Functions (EOF) efficiently and the global methodology (the SOM+EOF) is used to find the future values of a time series.

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

For the EOF, the advantages include high-dimensional linear projection of high-dimensional data without the decrease of dimensionality and the speed and the simplicity of the method.

The SOM+EOF includes the advantages of both individual methods, leading to a new accurate approximation methodology for the missing future values of a time series. The performance obtained in validation show the better accuracy of the new methodology.

It is also evident that the EOF is greatly dependent on a good initialization in order to produce accurate results. The SOM gives a good initialization even though the method alone is not so accurate. The two methods complete each other and work well together.

For further work, the modifications and performance upgrades of the global methodology are investigated and applied to other types of datasets and time series from other fields of science, for example climatology and finance.

Acknowledgment

Part the work of A. Sorjamaa and A. Lendasse is supported by the project of New Information Processing Principles, 44886, of the Academy of Finland. The work of A. Lendasse is supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

[1] T. Kohonen, *Self-Organizing Maps*. Springer-Verlag, Berlin, 1995.
[2] S. Wang, "Application of self-organising maps for data mining with incomplete data sets," *Neural Computing and Applications*, vol. 12, no. 1, pp. 42–48, 2003.
[3] R. Preisendorfer, *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, 1988.
[4] J. Boyd, E. Kennelly, and P. Pistek, "Estimation of eof expansion coefficients from incomplete data," *Deep Sea Research*, vol. 41, pp. 1479–1488, 1994.
[5] ESTSP2007 Conference: http://www.estsp2007.org.
[6] NN3 Competition: http://www.neural-forecasting-competition.com/index.htm.
[7] A. Sorjamaa and A. Lendasse, "Time series prediction using dirrec strategy." European Symposium on Artificial Neural Networks, ESANN 2006, Bruges (Belgium), 26-28 April, 2006, pp. 143–148.
[8] M. Cottrell and P. Letrémy, "Missing values: Processing with the kohonen algorithm." Applied Stochastic Models and Data Analysis, Brest, France, 17-20 May, 2005, pp. 489–496.
[9] SOM Toolbox: http://www.cis.hut.fi/projects/somtoolbox/.
[10] A. Lendasse, V. Wertz, and M. Verleysen, "Model selection with cross-validations and bootstraps - application to time series prediction with rbfn models," in *LNCS*, no. 2714, ICANN/ICONIP (2003). Berlin: Springer-Verlag, 2003, pp. 573–580.
[11] SOM+EOF Toolbox: http://www.cis.hut.fi/projects/tsp/?page=Downloads.