A prediction method for nonlinear time series analysis by combining the false nearest neighbors and subspace identification methods

I. Marín Carrión
University of Castilla-La Mancha
Applied Physics Department
Avd. España s/n, 02071 Albacete
SPAIN
e-mail: ismael.marin@uclm.es

E. Arias Antúnez
University of Castilla-La Mancha
Computing System Department
Avd. España s/n, 02071 Albacete
SPAIN
e-mail: enrique.arias@uclm.es

M. M. Artigao Castillo
University of Castilla-La Mancha
Applied Physics Department
Avd. España s/n, 02071 Albacete
SPAIN
e-mail: mariamar.artigao@uclm.es

J. J. Miralles Canals
University of Castilla-La Mancha
Applied Physics Department
Avd. España s/n, 02071 Albacete
SPAIN
e-mail: juan.miralles@uclm.es

***Abstract*—Nonlinear time series analysis is a powerful methodology that permits to predict the temporal evolution of some kinds of dynamical systems from characteristic quantities, such as the minimal embedding dimension or the maximal Lyapunov exponent. In fact, one of the most important goals of nonlinear analysis of experimental time series is the prediction. The subspace identification methods provides a good framework to model a system, both deterministic and stochastic, in a easily way. In order to make predictions, we propose a method which combine the minimal embedding dimension obtained by the method of false nearest neighbors and a model estimated by means of a subspace identification method. The results, in terms of predicted error, show the reliability of this new approach.**

***Index Terms*—Nonlinear Time Series Analysis, False Nearest Neighbors Method, Subspace Identification Methods, Prediction**

I. Introduction

Since the discovery of chaos, a growing interest in this field of research has risen rapidly. Nonlinear time series analysis is a powerful methodology that permits to predict the temporal evolution of some kinds of dynamical systems from characteristic quantities, such as the minimal embedding dimension or the maximal Lyapunov exponent. These characteristic quantities are extracted from study of time series obtained by any variable of the dynamical system. Thus, nonlinear time series analysis provides tools that bridge the gap between experimentally observed irregular behavior and deterministic chaos theory [5], [12], [14], [15].

Prediction is one of the main goals of the time series analysis. When a prediction of the behavior of the dynamical system has to be given, local and global predictors as those which appears in TISEAN [4], Artificial Neural Networks or others prediction methods based on ARIMA framework can be used.

On the other hand, the Subspace Identification Methods (SIMs) [18], [19] provides a good framework to model a system, both deterministic and stochastic, in a easily way. By using the model obtained by means of a SIM it is possible to predict the future behavior of the system. In this case, only one parameter is needed, that is, the order of the system.

In this paper we propose a new approach for the prediction of air temperature by combining the method of the False Nearest Neighbors (FNN) [6] and a SIM. The FNN method is used to calculate the minimal embedding dimension of a dynamical system.

This paper is organized as follows. Section II introduces the method of false nearest neighbors and the framework of the subspace system identification, and then a method which combines both sources of information is proposed. Section III presents the case studies and the experimental results. Finally, the conclusions and the future works are outlined in Section IV.

II. The method of False Nearest Neighbors and Subspace System Identification

This section introduces the method of false nearest neighbors and the framework of subspace identification methods, and then it proposes a method which combines both sources of information.

A. The method of False Nearest Neighbors

Dynamical systems are studied from two different view points. One is from an previously known model which explains

its behavior while the other is from a time series carried out by means of successive data acquisition per constant time periods $\{y_i\}_{i=1}^{i=N}$. This becomes the basis of nonlinear time series analysis. This methodology is based on the reconstruction of state space in a dynamical system from theorem of Takens [16]. The basic idea is the dynamical states space M of a dynamical system with dimension m is able to be characterized uniquely by m independent quantities. One of these systems of independent quantities are the own coordinates of the phases space (more precisely, the coordinates related to the basis that causes this phases space M).

$$y^{\rho}(t) = (y_1(t), y_2(t), \ldots, y_i(t), \ldots, y_m(t))^T \quad (1)$$

The most important phase space reconstruction technique is the method of delay. This technique takes m consecutive elements from the time series as coordinates in the phase space in order to find a vectorial space that contains the same information as the original states space [13]. This implies transforming a set of scalar data of dimension 1 (the time series) into vectorial data of dimension m (the reconstructed space phase). This phase space is the natural basis to formulate nonlinear time series algorithms from the chaos theory, rather than the time or the frequency domain. Vectors in this new space, the embedding space, are formed from time delayed values of the scalar measurements:

$$\vec{y_i} = [y_{i-(\mathbf{m}-1)\tau}, ..., y_{i-\tau}, y_i]^T \quad i : 1, 2, 3, ..., N-(m-1)\tau \quad (2)$$

In Eq. (2), the number m of elements is called the embedding dimension, and the time τ is generally referred to as the delay.

In practice, the natural questions are what time delay and what embedding dimension is the most appropriate for the reconstruction in the phase space. There are several methods to find out the minimal embedding dimension, m, i.e. global embedding dimension. A noteworthy method is the aforementioned method of false nearest neighbors. This method identifies the number of "false nearest neighbors", points that appear to be nearest neighbors because the embedding space is too small.

The FNN method supposes that the minimal embedding dimension for a time series $\{y_i\}$ is m_0. In this way, the reconstructed system in a m_0-dimensional delay space is a one-to-one image of the system in the original phase space. Thus, the neighbors of a given point are mapped onto neighbors in the delay space. If an $m-$dimensional space ($m < m_0$) is considered, then the topological structures are not preserved and the points are projected into neighborhoods of other points to which they would not belong in higher dimensions. In this case, these points are called *false neighbors*.

The idea of the FNN method is to measure the distances between a point $\vec{y_i}$ and its nearest neighbor $\vec{y_j}$; as this dimension increases, this distance should not change if the points are really nearest neighbors. If we define the distance between a point and its nearest neighbor using Euclidean distance, we can evaluate the change in distance by adding one more dimension and then we can look at the relative change in the distance as a way to see if our points were not really close together but a projection from a higher phase space. The criterion for falseness is thus

$$\frac{\mid y_{i+1} - y_{j+1} \mid}{\parallel \vec{y_i} - \vec{y_j} \parallel} > R_t, \quad (3)$$

where R_t is a threshold value. Using this criterion we can then test our sequence of points and, as dimension increases, find where the percentage of nearest neighbors goes to 0.

B. Subspace System Identification

Subspace-based system identification is a branch that has been recently developed in system identification (since beginning of 1990's), which has attracted much attention, owing to its computational simplicity and effectiveness in identifying dynamic state-space linear multivariable systems.

The three most well-known subspace algorithms are N4SID [17], MOESP [19] and CVA [7]. N4SID is one of the most popular classes of subspace algorithms. N4SID are part of the set of combined deterministic-stochastic subspace algorithms, but our interest is in the stochastic part because it computes state space models using only output data. So that, N4SID is the class of subspace algorithms that have been used for the proposed scheme of modeling and predicting. Following is described the N4SID class and then the stochastic identification.

Numerical algorithms for Subspace State Space System Identification

The greater part of the systems identification literature is concerned with computing polynomial models, which are however known to typically give rise to numerically ill-conditioned mathematical problems, especially for Multi-Input Multi-Output systems. Numerical algorithms for Subspace State Space System Identification (N4SID) are then viewed as the better alternatives [17]. This is especially true for high-order multivariable systems, for which it is not trivial to find a useful parameterization among all possible parametrizations. This parametrization is needed to start up the classical identification algorithms [8], which means that a priori knowledge of the order and of the observability (or controllability) indices is required.

With N4SID algorithms, most of this a priori parametrization can be avoided. Only the order of the system is needed and it can be determined through inspection of the dominant singular values of a matrix that is calculated during the identification. The state space matrices are not calculated in their canonical forms (with a minimal number of parameters), but as full state space matrices in a certain, almost optimally conditioned basis (this basis is uniquely determined, so that there is no problem of identifiability). This implies that the observability (or controllability) indices do not have to be known in advance.

Another major advantage is that N4SID algorithms are non-iterative, with no nonlinear optimization part involved. This

is why they do not suffer from the typical disadvantages of iterative algorithms, e.g. no guaranteed convergence, local minima of the objective criterion and sensitivity to initial estimates.

For classical identification, an extra parametrization of the initial state is needed when estimating a state space system from data measured on a plant with a non-zero initial condition.

Stochastic Identification

This subsection treats the subspace identification of purely stochastic systems with no external input ($u_k \equiv 0$). The stochastic identification problem thus consists of computing the stochastic system matrices from given output data only.

Stochastic subspace identification algorithms compute state space models from given output data. Following it states the stochastic (subspace) identification problem.

Given: s measurements of the output $y_k \in \mathbb{R}^l$ generated by the unknown stochastic system of order n:

$$x_{k+1}^s = Ax_k^s + w_k, \tag{4}$$

$$y_k = Cx_k^s + v_k, \tag{5}$$

with w_k and v_k zero mean, white vector sequences with covariance matrix:

$$\mathbf{E}[\begin{pmatrix} w_p \\ v_p \end{pmatrix} \begin{pmatrix} w_q^T & v_q^T \end{pmatrix}] = \begin{pmatrix} Q & S \\ S_T & R \end{pmatrix} \delta_{pq}. \tag{6}$$

Determine:

- The order n of the unknown system
- The system matrices $A \in \mathbb{R}^{n\times n}$, $C \in \mathbb{R}^{l\times n}$ up to within a similarly transformation and $Q \in \mathbb{R}^{n\times n}$, $S \in \mathbb{R}^{n\times l}$, $R \in \mathbb{R}^{l\times l}$ so that the second order statistics of the output of the model and of the given output are equal.

C. Combination of FNN and SIM methods

In the literature we can find three basic methods used to choose the minimal embedding dimension.

1) Computing an invariant on the attractor [3].
2) Singular Value Decomposition (SVD) [1].
3) The FNN method [6], which has been previously introduced.

The subspace identification methods provide a good framework to model a system, both deterministic and stochastic, in a easily way. By using the model obtained by means of subspace identification methods, it is possible to predict the behavior of the system. In this case, only one parameter is needed, that is, the order of the system.

The subspace identification methods are based on SVD information asking for an order of the system, which is used to generate a model of it. The SVD procedure and the FNN method are used to choose the minimal embedding dimension of a system. So that, it is possible to combine both information sources in order to automatically generate a model of the system. The minimal embedding dimension obtained by FNN method is passed to the subspace identification method to generate the model. This dimension is now considered the order of system.

A MATLAB program (called ed_n4sid) has been made using the $n4sid()$ and $predict()$ functions of the System Identification Toolbox of MATLAB. This program loads the time series y, estimate a model based on first $\frac{n}{2}$ data points using the $n4sid()$ function, and evaluate the $k-$step ahead predictions on the second half of y. The order used to estimate the model is the minimal embedding dimension m of the time series, instead of using SVD.

Prediction based on a model means forecasting the model response $k-$steps ahead into the future using the current and past values of measured input and output. k, or kTs time units where Ts is the sampling interval, is the prediction horizon. To predict the model's response $k-$steps into the future from the current time t, one needs to know inputs up to time $t+k$ and outputs up to time t:

$$\begin{aligned}\hat{y}(t+k) = f(&u(t+k), u(t+k-1), ..., \\ &u(t), u(t-1), ..., u(0), \\ &y(t), y(t-1), y(t-2), ..., y(0))\end{aligned}$$

$u(0)$ and $y(0)$ are the *initial states*. $f()$ represents the *predictor* whose form depends on the model structure. However, ed_n4sid method is based on stochastic subspace system identification, and therefore no inputs are considered.

The model is evaluated in state space form, and the state equations are simulated $k-$steps ahead with initial value $y(t-k) = \hat{y}(t-k)$, where $\hat{y}(t-k)$ is the Kalman filter state estimate.

Pseudocode 1 describes the proposed method ed_n4sid.

Algorithm 1 ed_n4sid method

Program $ed\_n4sid : (y, m, k)$
Inputs: y is a time series of n data points; m is the minimal embedding dimension of the system (order of the system); and k is the prediction horizon
Outputs: $\hat{y}$ is the predicted output; and the Best Fit (BF) metric

1: Loading time series of n length
2: Estimating a state-space model using the $n4sid()$ function with the first $\frac{n}{2}$ points of the y. The order used is the minimal embedding dimension m
3: Predicting the output ($\hat{y}$) $k-$steps ahead using the $predict()$ function. This function takes the estimated model and the second half of y
4: Plotting the second half of y and $\hat{y}$
5: Computing the best fit of the predicted output $\hat{y}$

This method uses the *Best Fit* (BF) metric as prediction error. The method displays the percentage of the output that the model reproduces (best fit), computed using Eq. (7):

$$Best\ Fit = \left(1 - \frac{|y - \hat{y}|}{y - \bar{y}}\right) \times 100 \tag{7}$$

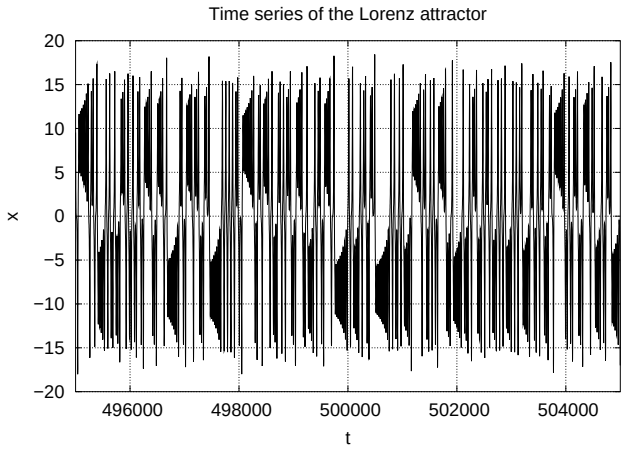


Fig. 1. Time series of the Lorenz attractor.

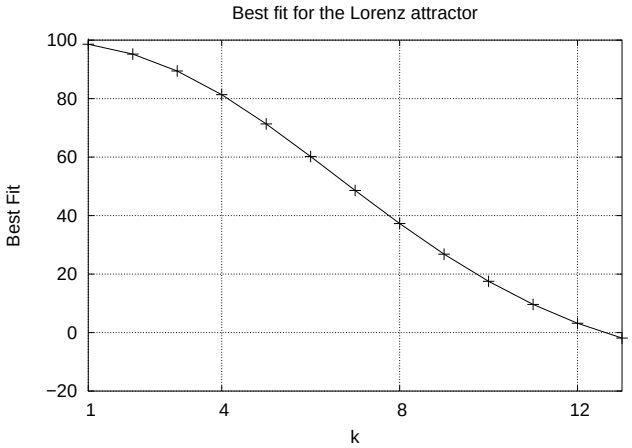


Fig. 2. Best Fit for the Lorenz attractor.

In Eq. (7), y is the measured output, $\hat{y}$ is the predicted model output, and $\bar{y}$ is the mean of y. 100% corresponds to a perfect fit, and 0% indicates that the fit is no better than guessing the output to be a constant ($\hat{y} = \bar{y}$). Because of the definition of best fit, it is possible for this value to be negative.

III. Case Studies and Experimental Results

This section describes the case studies selected, and the results of appliying the proposed prediction method in each case. The Lorenz attractor, which is a classical example of continuous-time chaotic system, an ECG signal and a temperature time series form the case studies.

A. The Lorenz Attractor

The Lorenz system [9] shows how the state of a dynamical system (the three variables of a three-dimensional system) evolves over time in a complex, non-repeating pattern, which is often described as beautiful. The equations that describe the system were introduced by E. Lorenz in 1963, who derived it from the simplified equations of convection rolls arising in the equations of the atmosphere. These equations are the following

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y - x) \\ \frac{dy}{dt} &= x(r - z) - y \\ \frac{dz}{dt} &= xy - bz, \end{aligned} \tag{8}$$

where σ, r and b are the parameters of the Lorenz system. σ is called the *Prandtl number*, r is called the *Rayleigh number* and b is a *geometric factor*.

Fig. 1 shows a part of the Lorenz time series (the transitory part has been removed).

The minimal embedding dimension of this case is 3, which has been computed by using the TISEAN routines. As it is commented in Section II, this dimension is now considered the order of system, which is passed to the subspace identification method to generate the model and then to predict k steps ahead.

Fig. 2 shows the prediction error in function of k according to best fit metric for the Lorenz attractor.

The accuracy of the method for the Lorenz attractor is greater than 50% for $k \leq 6$, decreasing the accuracy until $k = 12$.

Fig. 3 shows a part of the original and predicted output using two different values of k for the Lorenz case study.

The one−step ahead prediction for the Lorenz attractor (red) and the original time series (blue) are practically equal, while the $5-$steps ahead prediction shows a prediction quite similar to the original time series with small differences. In fact, the $5-$steps ahead prediction has around $BF = 71\%$.

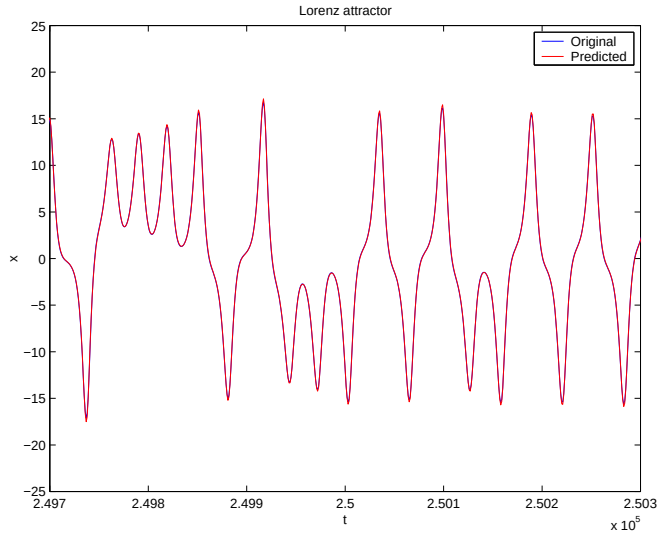
B. ECG Signal

ElectroCardioGraphy (ECG) is a transthoracic interpretation of the electrical activity of the heart over time captured and externally recorded by skin electrodes. It is a noninvasive recording produced by an electrocardiographic device.

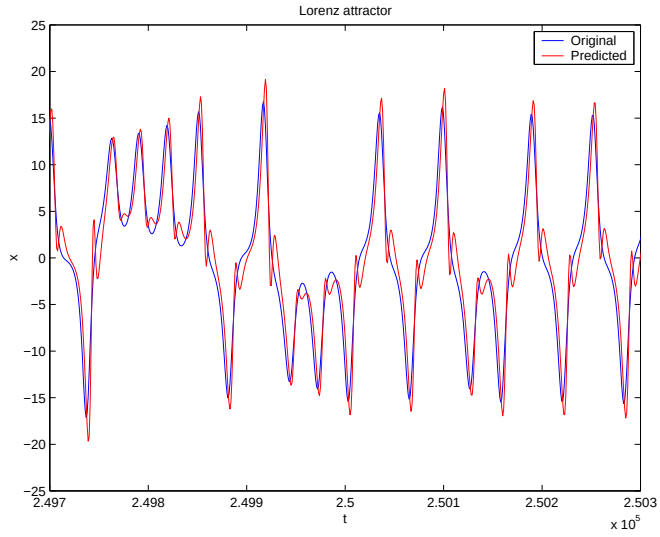
Electrical impulses in the heart originate in the sinoatrial node and travel through the intimate conducting system to the heart muscle. The impulses stimulate the myocardial muscle fibers to contract and thus induce systole. The electrical waves can be measured at electrodes placed at specific points on the skin. Electrodes on different sides of the heart measure the activity of different parts of the heart muscle. An ECG displays the voltage between pairs of these electrodes, and the muscle activity that they measure, from different directions, can also be understood as vectors. This display indicates the overall rhythm of the heart and any weaknesses in different parts of the heart muscle.

The ECG time series has been simulated by ECGSYN [2], [10], [11]. The ECG sampling frequency is 256Hz. Simulation is 60bpm of heart rate mean for a healthy person. Standard deviation of heart rate is 1bpm.

Fig. 4 shows the ECG time series.



(a)



(b)

Fig. 3. Original and predicted output for the Lorenz attractor. (a) $k = 1$. (b) $k = 5$.

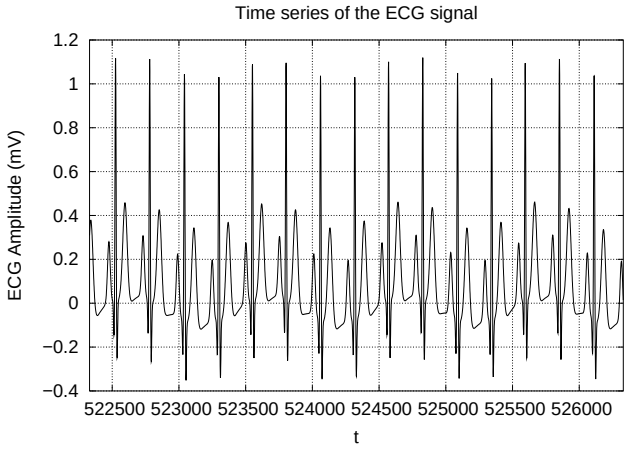


Fig. 4. Time series of the ECG.

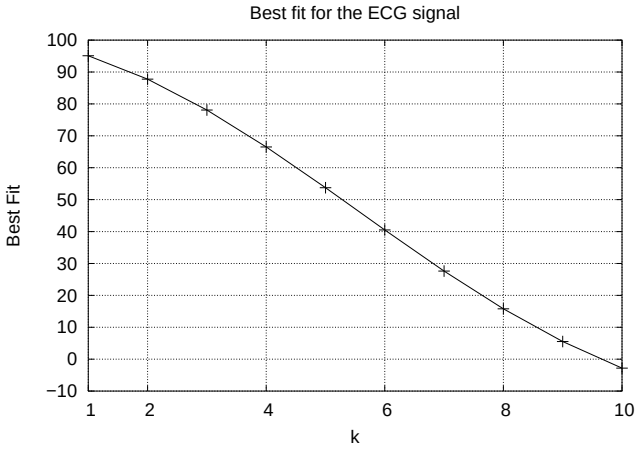


Fig. 5. Best Fit for the ECG signal.

The value of the embedding dimension is 3, according to the output of the FNN method.

Fig. 5 shows the prediction error in function of k according to best fit metric for the ECG signal.

The accuracy of the method for the ECG signal is greater than 50% for $k \leq 5$, decreasing until $k = 9$.

Fig. 6 shows a part of the original and predicted output using two different values of k for the ECG signal.

As in the previous case, the prediction for the ECG signal is quite similar to the original time series. Using the $3-$steps ahead prediction, with around $BF = 78\%$, the QRS complex and P- and T-waves of the ECG signal can be seen.

C. Air Temperature Data

We have also considered a one-year time series of air temperature data. The weather data have been recorded by means of a meteorological station (see Fig. 7) located on the top of the *Escuela Politécnica Superior de Albacete* building, University of Castilla-La Mancha. This building is placed in the urban area of Albacete (Spain). The meteorological station is property of the "Interdisciplinary Research Group in Dynamical Systems" (IRGDS). This case study has special relevance in Castilla-La Mancha due to the fact that the importance of the meteorology/climatology in the economy of this region, in the sense that one of the most important economic sectors in Castilla-La Mancha consist on agriculture.

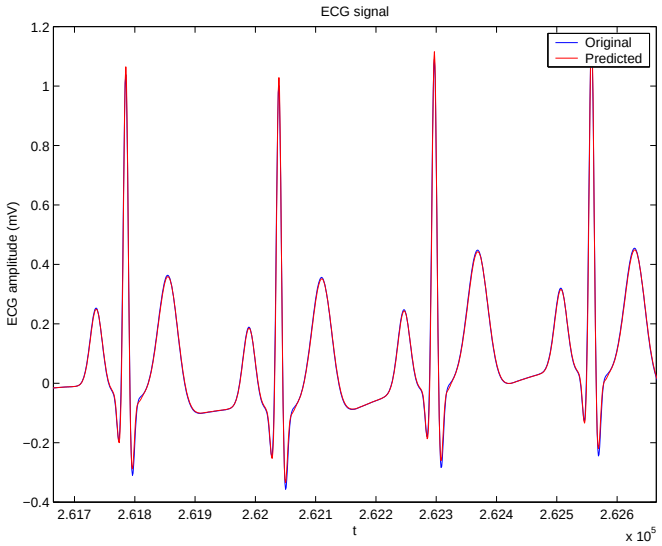
Fig. 7 shows a picture of the meteorological station (left) and a picture of the temperature sensor (right).

Fig. 8 shows the time series of temperature data.

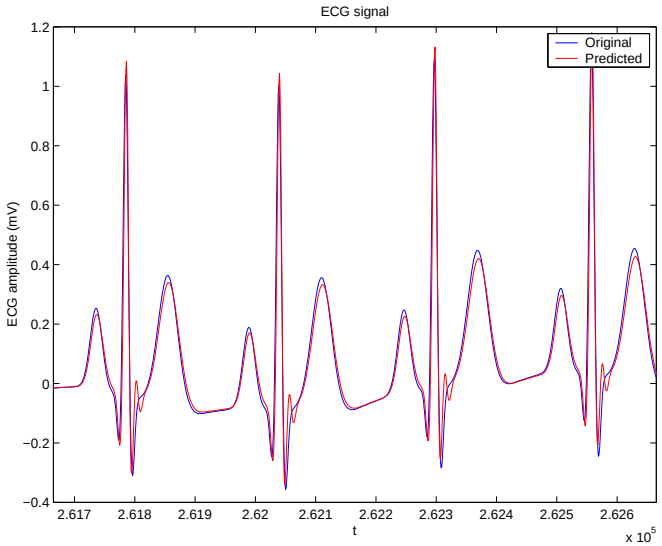
The minimal embedding dimension of this case is 3. Then, this is passed to the subspace identification method to generate the model and then to predict k steps ahead.

Fig. 9 shows the prediction error in function of k according to best fit metric for the time series of temperature.

The time series of temperature shows a good accuracy for $k \leq 30$ with around $BF = 55\%$, decreasing rapidly until $k = 40$.



(a)



(b)

Fig. 6. Original and predicted output for the ECG signal. (a) $k = 1$. (b) $k = 3$.

(a)

(b)

Fig. 7. Picture of meteorological station. (a) General view of the meteorological station. (b) View of the temperature sensor.

Moreover, Fig. 10 shows a part of the original and predicted output using two different values of k for the temperature time series. It shows that the prediction follows the maximum and minimum of temperature data during the presented interval of time, using a $15-$ahead predictions (around $BF = 88$). The one$-$step ahead prediction shows a fit really similar to the original time series.

IV. Conclusions and Future Work

This paper has proposed a method for modeling and predicting with time series data, which is based on subspace system identification. N4SID is the class of subspace algorithms chosen for this work. The subspace identification methods are based on SVD information asking for an order of the system, which is used to generate a model of it. The SVD procedure and the FNN method are used to choose the minimal

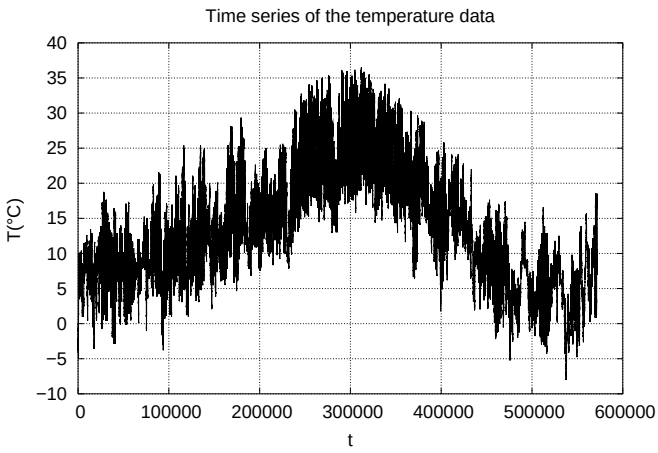


Fig. 8. Time series of the temperature data.

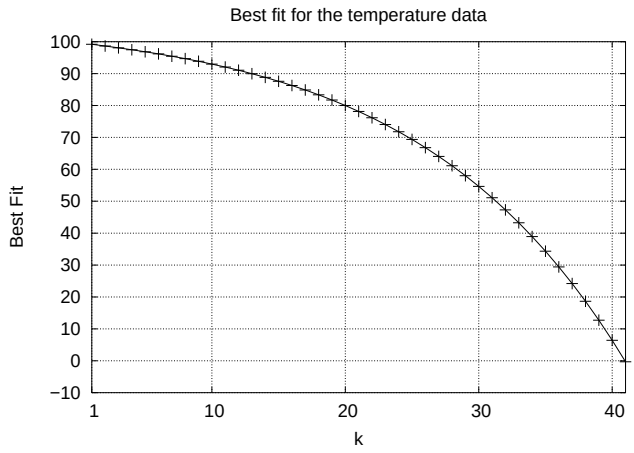


Fig. 9. Best Fit for the temperature time series.

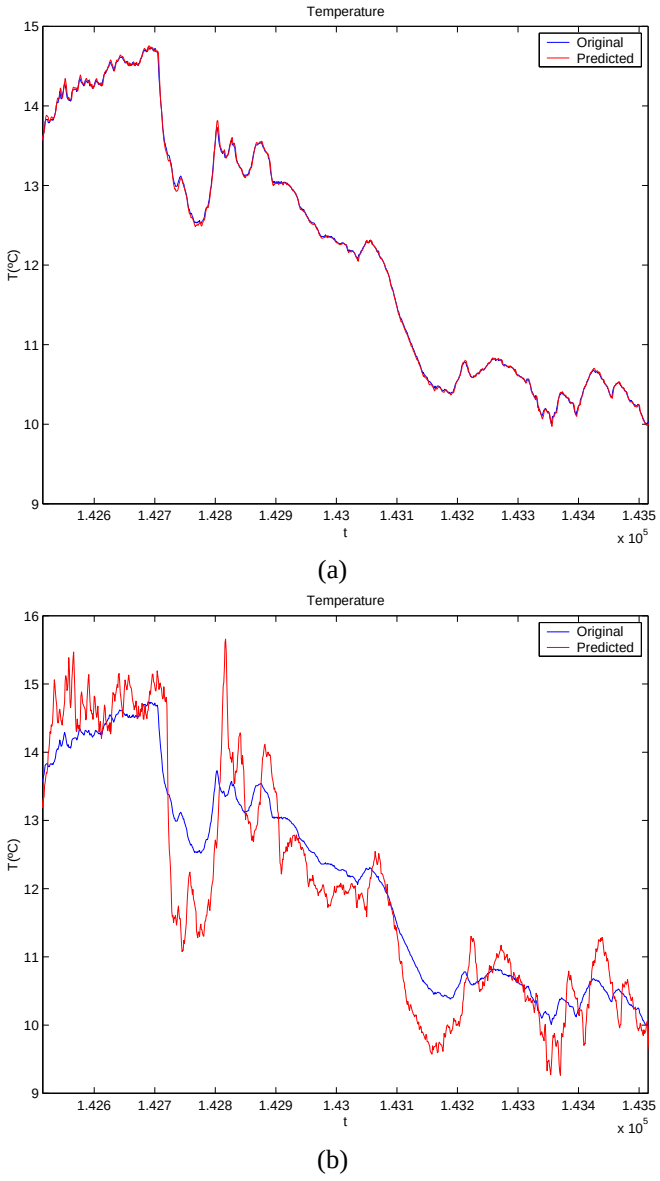


Fig. 10. Original and predicted output for the temperature time series. (a) $k = 1$. (b) $k = 15$.

embedding dimension of a system, which unfolds the attractor in the projected state space with no overlaps. So that, it is possible to combine both information sources in order to automatically generate a model of the system. The minimal embedding dimension obtained by FNN method is passed to the subspace identification method to generate the model. This dimension is now considered the order of system. The model is evaluated in state-space form, and the state equations are simulated k steps ahead. The paper presents some results that corroborates the goodness of the proposed method.

Although the ed_n4sid method has achieved good results, it is interesting to carry out an more exhaustive analysis of the method and make an comparative study of other relevant methods in this area. Moreover, it is expected to asses the accuracy of the method for hole filling of time series.

Finally, it can be extended to other time series models, such as ARIMA framework.

ACKNOWLEDGMENTS

This work has been partially supported by the Spanish CICYT project CGL2008-05688-C02-01/CLI and by the Castilla-La Mancha project JCCM-PCI-05-019.

REFERENCES

[1] D. S. Broomhead and G. P. King, "Extracting qualitative dynamics from experimental data", *Phys. D*, vol. 20, pp. 217–236, 1986.
[2] ECGSYN: http://www.physionet.org/physiotools/ecgsyn/.
[3] P. Grassberger and I. Procaccia, "Characterization of strange attractors", *Phys. Rev. Lett.*, vol. 50, pp. 346–349, 1983.
[4] R. Hegger, H. Kantz, and T. Schreiber, "Practical implementation of nonlinear time series methods: The TISEAN package", *Chaos*, vol. 9, pp. 413–435, 1999.
[5] D. T. Kaplan and L. Glass, *Understanding Nonlinear Dynamics*. New York: Springer, 1995.
[6] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction", *Phys. Rev. A*, vol. 45, pp. 3403–3411, 1992.
[7] W. E. Larimore, "Canonical variate analysis in identification, filtering, and adaptive control", in *Proc. 29th IEEE Conf. Decis. Control.*, Honolulu, 1990, pp. 596–604.
[8] L. Ljung, *System Identification: Theory for the User*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
[9] E. Lorenz, "Deterministic nonperiodic flow", *J. Atmos. Sci.*, vol. 20, pp. 130–141, 1963.
[10] P. E. McSharry, G .D. Clifford, L. Tarassenko, and L. A. Smith, "A dynamical model for generating synthetic electrocardiogram signals", *IEEE Trans. Biomed. Eng.*, vol. 50, pp. 289–294, 2003.
[11] P. E. McSharry and G. D. Clifford, "Open-source software for generating electrocardiogram signals", in *Proc. 3rd IASTED Int. Conf. Biomed. Eng.*, Innsbruck, 2005.
[12] E. Ott, *Chaos in Dynamical Systems*. Cambridge: Cambridge University Press, 1993.
[13] N. H. Packard, J. P. Crutchfield, J .D. Farmer, and R. S. Shaw, "Geometry from a time series," *Phys. Rev. Lett.*, vol. 45, pp. 712–716, 1980.
[14] H. G. Schuster, *Deterministic Chaos*. Weinheim: VCH, 1989.
[15] S. H. Strogatz, *Nonlinear Dynamics and Chaos*. Reading, MA: Addison-Wesley, 1994.
[16] F. Takens, "Detecting strange attractors in turbulence", in *Dynamical Systems and Turbulence, Warwick 1980*, D.A. Rand and L-S. Young, Eds. New York: Springer, 1981, pp. 366–381.
[17] P. Van Overschee and B. De Moor, "N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems", *Autom.*, vol. 30, pp. 75–93, 1994.
[18] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems: Theory Implementation Applications*. Dordrecht: Kluwer Academic Publishers, 1996.

[19] M. Verhaegen and P. Dewilde, “Subspace identification, part i: The output-error state space model identification class of algorithms”, *Int. J. Control.*, vol. 56, pp. 1187–1210, 1992.