An innovative hybrid neuro-wavelet method for reconstruction of missing data in astronomical photometric surveys

Giacomo Capizzi[1], Christian Napoli [2,*], and Lucio Paternó[3]

[1]Dpt. of Electric, Electronic and Informatics Engineering, University of Catania, Italy
[2]Dpt. of Mathematics and Computer Science, University of Catania, Italy
[3]National Institute for Astrophysics (INAF), Italy

The investigation of solar-like oscillations for probing the star interiors has encountered a tremendous growth in the last decade. For ground based observations the most important difficulties in properly identifying the true oscillation frequencies of the stars are produced by the gaps in the observation time-series and the presence of atmospheric plus the intrinsic stellar granulation noise, unavoidable also in the case of space observations. In this paper an innovative neuro-wavelet method for the reconstruction of missing data from photometric signals is presented. The prediction of missing data was done by using a composite neuro-wavelet reconstruction system composed by two neural networks separately trained. The combination of these two neural networks obtains a "forward and backward" reconstruction. This technique was able to provide reconstructed data with an error greatly lower than the absolute a priori measurement error. The reconstructed signal frequency spectrum matched the expected spectrum with high accuracy.

***Index Terms*—Kepler Mission, Recurrent Neural Networks, Wavelet Theory, Photometry, Missing Data**

I. Introduction

The investigation of solar-like oscillations for main sequence, sub giant and red giant stars for probing the star interiors has encountered a tremendous growth in the last decade. This science, known as Asteroseismology, is fairly increasing our knowledge about stellar physics, especially after the launch of the NASA space mission Kepler in 2009 [1].

The data acquired for the study of solar-like oscillations are mainly of two different types, spectroscopic and photometric. Although in both cases they are in the form of a temporal sequence of measurements (time-series) and are able to probe the same physical quantities, the information carried on is not equivalent and their usage appears to be complementary. Ground-based observations, which usually detect oscillations by exploiting very high-precision Doppler shift measurements of the spectroscopic lines, can probe a wider number of modes of oscillations because of their high sensitivity to spatial resolution upon the stellar disk. Nonetheless, ground-based projects are able to follow up one target per time and heavily suffer the alternating of day and night due to Earths rotation, which hence does not allow for continuous-time observations. Furthermore, the effort and workload required for assembling the spectrometers used to acquire the data do not allow to use such systems on space.

For ground-based observations the most important difficulties in properly identifying the true oscillation frequencies of the stars are produced by the gaps in the observation time-series and the presence of atmospheric plus the intrinsic stellar granulation noise, the latter unavoidable also in the case of space observations. The gaps are caused by the alternation of day and night and casual interruptions of data flow due to bad weather conditions; the first introduces possible shifts of 11.57 μHz n the identified frequencies and the second spurious frequencies. The noise can produce peaks whose amplitude is even larger than the real stellar frequencies. All the mentioned disturbs make the identification of stellar oscillations uncertain in several cases. In this paper the problem of data prediction in order to reconstruct the gaps present in the observation time-series has been addressed by using an hybrid computation methods based on wavelet decomposition and recurrent neural networks (RNNs).

Wavelet analysis has been used in order to reduce the data redundancies and selectively remove stellar granulation noise so obtaining a representation that can express their intrinsic structure, while the neural networks (NNs) are used for the exploiting the complexity of non-linear data correlation and to perform the data prediction. In order to minimize the error propagation, we designed a composite network, with doubled neural paths, to obtain a "forward and backward" reconstruction. This composite WRNN uses as input several time steps of the signal, in the past and in the future with respect to the gap.

II. Kepler data as a probe for testing WRNN method

New missions based on photometric acquisitions have been launched on space in the few past years. The latest one in particular, the NASA Kepler mission, which is presently in the middle of its running, is providing an enormous amount of an unprecedented quality data, with a combined differential photometric precision high to $2 \cdot 10^6$ for a 12^{th} magnitude solar-like star for a 6.5 hour integration [2]. In fact, the photometric observations allow the great advantage of acquiring brightness measurements on hundred of targets at the same time and, most of all, they can be carried out directly from

*Email: napoli@dmi.unict.it.
PUBLISHED ON: **11th International Conference, ICAISC 2012, Zakopane, Poland, April 29-May 3, 2012, Proceedings, pp. 21-29, 2012**

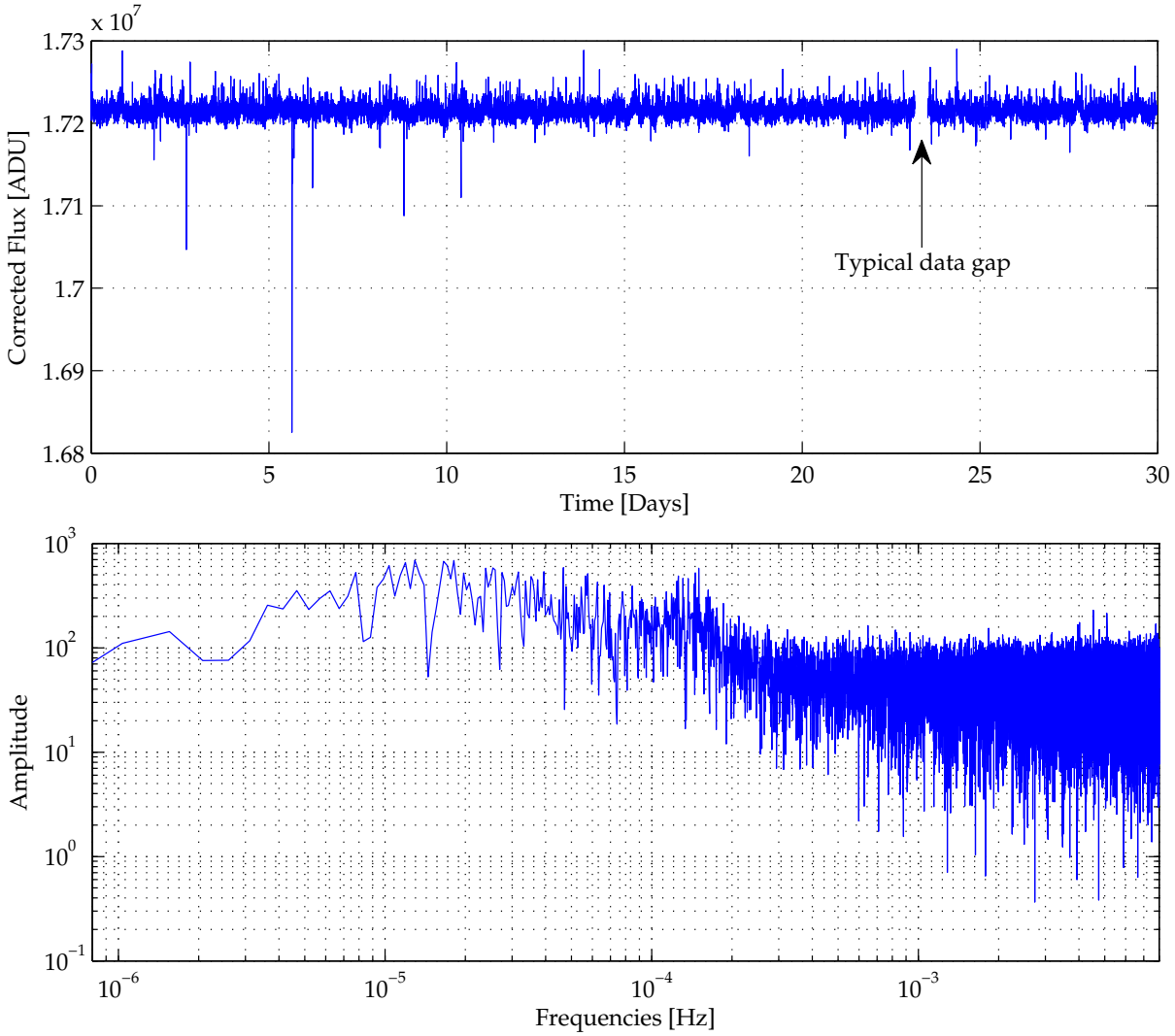

Fig. 1. The light flux time-series and the relative power spectrum

space, allowing scientists to weed out the problem of the daily gap, which strongly hampers the quality of the results, but not the problem of granulation noise. Long term acquisitions of brightness variations on the surface of stars are able to tell us a lot about solar-like oscillations as they are directly correlated to variations in temperature of the surface layers. By the continuous production of new data sets, many interesting studies can be made upon the stars falling in the Kepler field o view (FOV), from early main sequence to red giants stars.

The data used in this paper were collected by Kepler satellite with a sampling rate of about 58.7 s, as light flux measurement and corrected flux estimation with the related absolute error. The data were relative to the star KIC 3102411 measured at short cadence in the season Q2.2. The most common way to analyze a time-series and thus to derive the frequencies of oscillations is to convert the data acquired on the time domain to a set of values that range in a frequency domain (the Power Spectrum). This is done in general by adopting a Fourier analysis on the time-series, both of radial velocities (from Doppler shift measurements) and of radiation flux counts (from photometric acquisitions) [3].

The result is shown in Fig. 1, where the upper panel represents the light-curve, i.e. a time-series of radiation flux counts, for a star observed by Kepler and the lower panel is its relative Fourier transform reported in a logarithmic scale. As clearly visible, a bump of power arises around 150 μHz, showing the typical pattern for a set of p-mode frequencies that roughly follows a gaussian shape peaked at its maximum frequency $\nu_{\max}$. As one can intuitively understand, the longer is the observation run, the higher is the frequency resolution at which the frequency peaks in the power spectrum can be measured. The presence of huge gaps equally spaced in the time-series, as in the case of the daily gap, causes the arising of fictitious peaks in the power spectrum, which are not real frequencies of oscillation and that consequently affect the identification of the true p modes by hampering the true pattern of the solar-like excess of power in the power spectrum.

III. The WRNN methodology

The reconstruction of missing data from photometric time-series was done by using a composite neuro-wavelet reconstruction technique. RNNs are able to predict the continuation

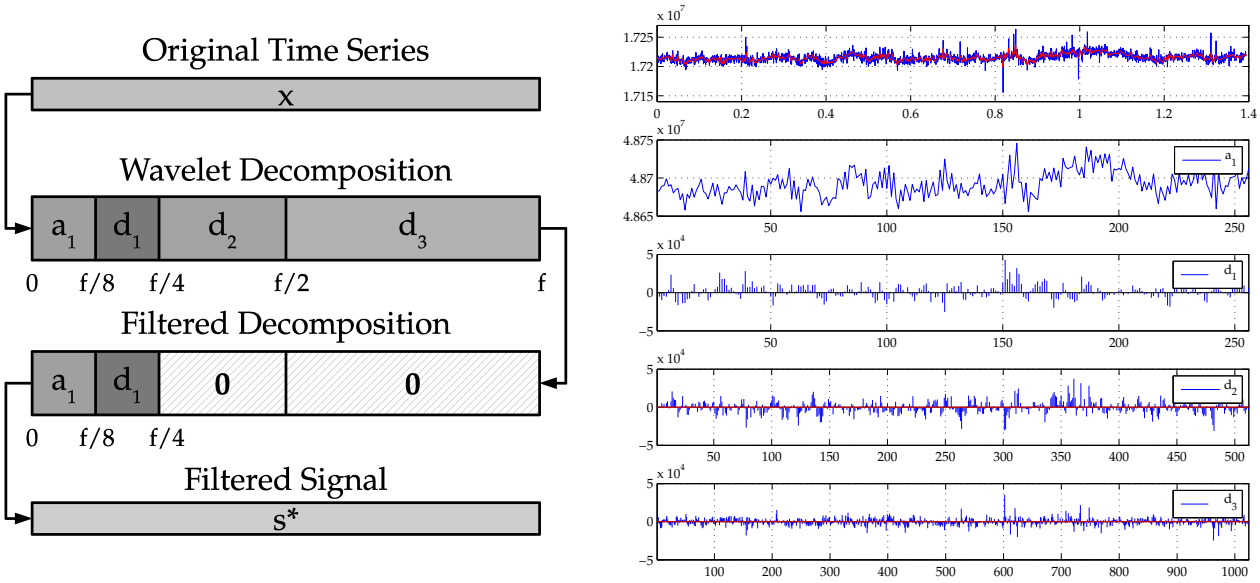

Fig. 2. Wavelet decomposition and thresholding

of a time series amounts to picking one of a class of functions so as to approximate the input-output behavior in the most appropriate manner. For deterministic dynamical behaviors, the observation at a current time point can be modeled as a function of a certain number of preceding observations. In such cases, the model used should have some internal memory to store and update context information [4], [5]. This is achieved by feeding the network with a delayed version of the past observations, commonly referred to as a delay vector or tapped delay line. These networks do not try to achieve credit assignment back through time but instead use the previous state as a part of the current input. Such a simple approach may be seen as a natural extension to feedforward the networks in much the same way that ARMA models generalize autoregressive models.

A network with a rich representation of past outputs, is a fully connected recurrent neural network, known as the Williams-Zipser network (NARX networks) [6], [7], [8]. For stochastic phenomena, like the considered ones, real time recurrent learning (RTRL) has proven to be very effective, in fact RTRL based training of the RNN is made upon minimizing the instantaneous squared error at the output of the first neuron of the RNN [6], [9]. The reconstruction system is composed by two NARX RNNs with the same topology and number of neurons but separately trained with RTRL algorithm (Fig. 4). A complete description of RTRL algorithm, NARX and RNNs can be found in [8]. The first one is trained to predict the signal samples one step ahead in the future, while the second one is trained to predict the signal samples one step backward in the past. The combination of these two neural networks obtains a "forward and backward" reconstruction (Fig. 4). This reconstruction technique was able to minimize the error propagation and, also, the possibility to conduct a double check verification of the reconstructed data.

At a first time the selected neural networks were trained to reconstruct missing data from a photometric time-serie which was yet proven to have an high cross-correlation degree. Although different kinds of topology and size variations were implemented, the system was not able to provide predictions with enough accuracy. On the other hand, there was evidence of misleading data sequences avoiding a correct training of the networks. At a successive step the same procedure was adopted, but, this time, providing as input the wavelet decompositions of the signal. A function $\psi \in L^2(\mathbb{R})$ that exhibit the following properties:

$$\int_{-\infty}^{+\infty} \psi(x)dx = 0 \tag{1}$$

$$||\psi(x)||^2 = \int_{-\infty}^{+\infty} \psi(x)\psi^*(x)dx = 1 \tag{2}$$

is called wavelet if it can be used to define a Hilbert basis, that is a complete system, for the Hilbert space $L^2(\mathbb{R}$ of square integrable functions. The Hilbert basis is constructed as the family of functions $\{\psi_{j,k} : j,k \in \mathbb{Z}\}$ by means of dyadic translations and dilations of $\psi, \psi_{j,k}(x) = \sqrt{2^j}(2^j x - k)$. For an extended treatment one can consult [10], [11], [12]. It is known in literature that the star granulation produces temporal variations in the light flux. These variations, at frequencies greater than 100 μHz, produce a quasi-white signal-related noise effect. Even if it is not possible to adapt a neural network to this kind of noise, the used wavelet decomposition permits to locate the coefficients bands related to frequencies from about 4250 μHz to higher frequencies. Thresholding to zeros these two related bands (Fig. 2) the resulting coefficients and residuals carries the relevant information for the predictions.

These wavelet coefficients were so provided as input ($\mathbf{u}_i(t)$) to the system. Another positive effect is that these wavelet

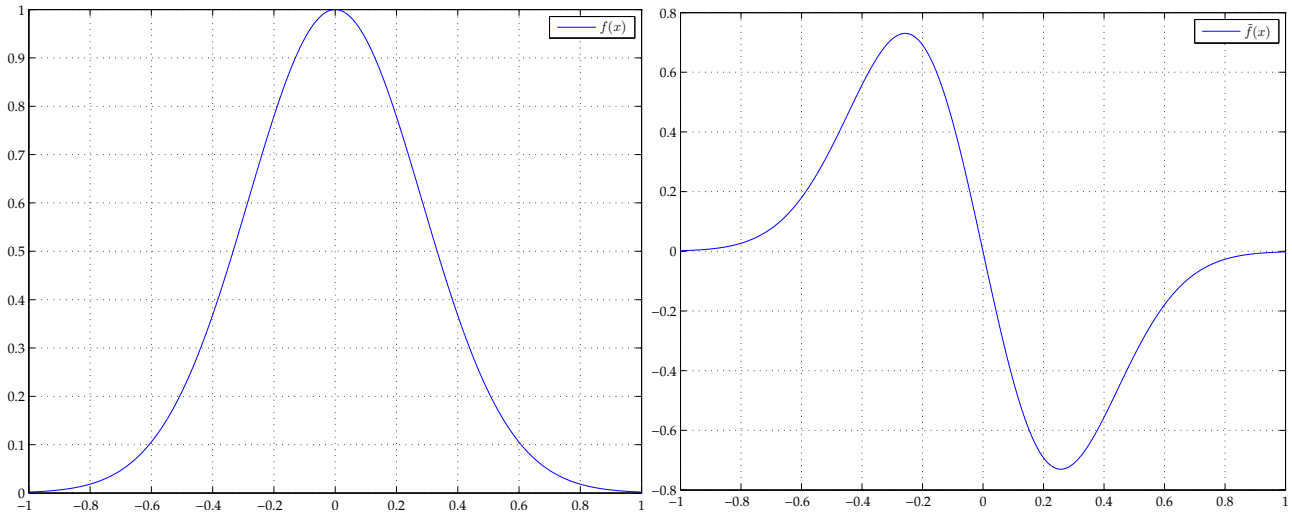

Fig. 3. The radial basis transfer function $f(x)$ and the relative wavelet function $\tilde{f}(x)$

coefficients provide a less redundant representation of the information carried out by the signal. This effect was proven to be an advantage for a correct and efficient training of recurrent neural networks [13], [14], [15]. As yet shown by a previous work of the authors [16], a properly designed hybrid neuro-wavelet recurrent network is able to execute wavelet reconstruction and prediction of a signal. The selected neural networks are composed by two hidden layers of 16 neurons and a single output neuron.

The wavelet decomposition of the time series is given as $N \times 4$ input vectors with a 3-step delay and a 1-step output feedback. While the forward network was trained with coefficients at time t0 to predict output signals $s(t_0 + 1)$, the backward network was trained to backward reconstruct the signal at a previous time $(t_0 - 1)$ in the past. For clarity, on describing the forward neural network with a functional of the type $F[\mathbf{u}(t)] = y(t+1)$, it follows that the backward network will be described by a similar $\tilde{F}[\mathbf{u}(-t)] = \tilde{y}(-t-1)$. In this manner, at the end of a correct training of the selected neural networks, it will be possible to reconstruct missing part of the data series using both the neural networks. The forward network will reconstruct forward in time the missing part from the beginning to his mid-point. The backward network will reconstruct backward in time from the ending to the same mid-point of the same gap. The resulting reconstructed signal $\tilde{s}$, from a signal s with missing data in the interval $[t_1; t_3]$ is

$$\tilde{s}(t) = \begin{cases} s(t) & t \in]-\infty, t_1[\\ y(t) & t \in [t_1, t_2[\\ \tilde{y}(t) & t \in]t_2, t_3] \\ s(t) & t \in]t_3, +\infty[\end{cases} \tag{3}$$

In the present paper the implemented WRNN is able to reconstruct a signal from wavelet coefficients, but it is also capable to predict these wavelet coefficients, and, then, to reconstruct the predicted signal. To obtain this behavior some rules had to be applied during the design and implementation work. For reasons that will be cleared ahead, all the hidden layers have a pair neuron number, and, also, to permit in sequence the wavelet coefficient exploitation and the signal reconstruction, a double hidden layer is required in the proposed architecture. As for the hidden layers the neurons activation function (transfer function) have to simulate a wavelet function. It is not possible to implement a wavelet function itself as transfer function for a forecast oriented time predictive neural network, this because wavelets do not verify some basic properties such as the absence of local minima, and does not provide by itself a sufficiently graded response [17].

In the existent range of possible transfer functions only some particular classes approximate the functional form of a wavelet. In this work the radial basis functions (radbas) were chosen as transfer functions, indeed this particular kind of functions well describes in first approximation half of a wavelet, even if these functions do not verify the properties shown by (1) and (2). Anyway, after scaling, shift and repetition of the chosen activation function, it is possible to obtain several mother wavelet filters. Let $f : [-1; 1] \to \mathbb{R}^+$ to be the choosen transfer function, then

$$\tilde{f}(x) = \tilde{f}(x + 2k) = \begin{cases} +f(2x+1) & x \in [-1, 0] \\ -f(2x-1) & x \in [0, 1] \end{cases} \tag{4}$$

verifies all the properties of a wavelet function. So it is possible for the selected neural networks to simulate a wavelet by using the radbas function defined in the $[1; 1]$ real domain. It is indeed possible to verify that

$$\int_{2h+1}^{2k+1} \tilde{f}(x)dx = 0 \quad \forall\ h < k \in \mathbb{Z} \tag{5}$$

It was shown that, in order to simulate a wavelet function, the chosen transfer functions must be symmetrically periodical to emulate a wavelet. This is the reason for choosing a pair number of neurons in the aim to have the same number of positive and negative layer weights in the reconstruction layer. Theoretically, if this happens, then the neuron pairs of the second layer are emulating exactly a reconstruction filter. Although this was a theoretical schema, there are strong

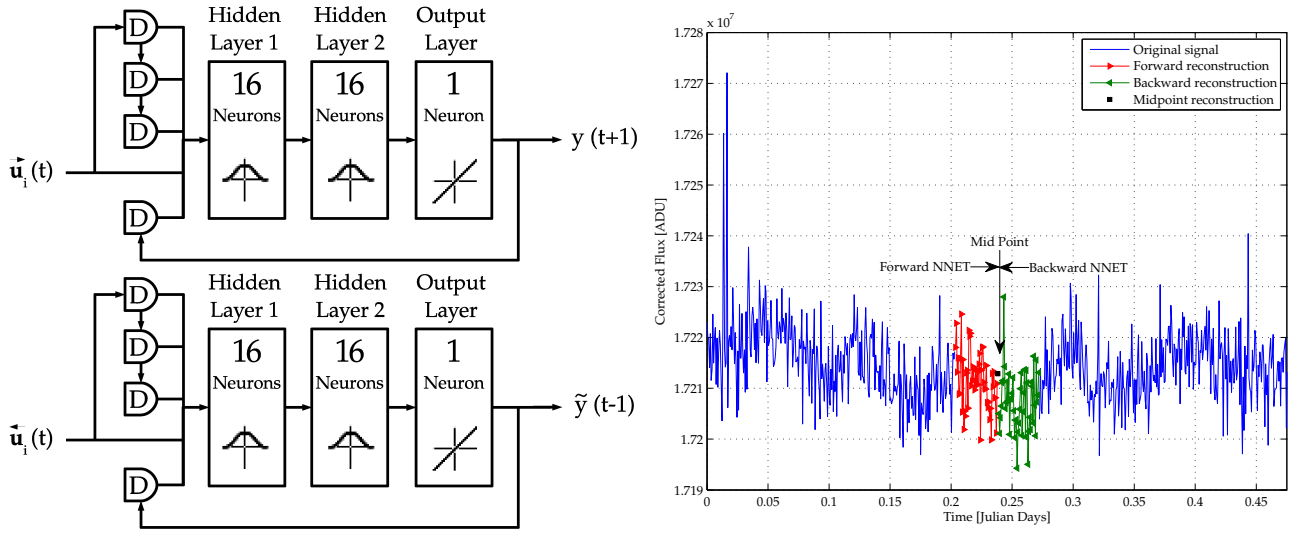

Fig. 4. Neural networks structures (left), Forward and Backward reconstruction (right)

reasons for the weights, in this experimental setup, to have a non-zero sum, because the neural network beyond to perform the inverse wavelet transform must perform also the signal prediction.

IV. Results and Conclusion

We performed simulations on one month photometric survey of the star KIC 3102411 observed during the season Q2.2 from the Kepler orbital telescope with a sampling rate of about 58.847 s and so a sampling frequency of almost $1.7 \cdot 10^2$ Hz. Wavelet analysis was used in order to remove the data sparsity and to thresh- old the higher frequencies (mostly characteristic of the star granulation and intrinsically affected by a signal-correlated time-evolving noise). In particular the lower two sub-bands of the decomposition were substituted with zero-vectors. In this manner the filtered reconstructed signal was transferred to the neural networks. To test the capabilities of the system, several gaps, ranging from 2 to 10 samples, were artificially placed at random positions in the data series.

The trained forward and backward reconstruction system was able to reconstruct the missing data with an error greatly lower than the absolute a priori measurement error. The reconstructed signal frequency spectrum matches the expected spectrum with high accuracy, as shown in Figs. 5 and 6. This paper has outlined the advantage of a composite hybrid neuro-wavelet system as advanced reconstruction tool for photometric time-series. This technique leads to implement a new generation of tools based on recurrent neural networks with the future possibility of further developments such as embedded system for data reconstruction of corrupted time-series for noise-affected survey contests.

Acknowledgment

This paper has been published in the final and reviewed version on **11th International Conference, ICAISC 2012, Zakopane, Poland, April 29-May 3, 2012, Proceedings, pp. 21-29, 2012** [18].

References

[1] W. J. Chaplin, H. Kjeldsen, J. Christensen-Dalsgaard, S. Basu, A. Miglio, T. Appourchaux, T. R. Bedding, Y. Elsworth, R. García, R. L. Gilliland *et al.*, "Ensemble asteroseismology of solar-type stars with the nasa kepler mission," *Science*, vol. 332, no. 6026, pp. 213–216, 2011.

[2] D. G. Koch, W. J. Borucki, G. Basri, N. M. Batalha, T. M. Brown, D. Caldwell, J. Christensen-Dalsgaard, W. D. Cochran, E. DeVore, E. W. Dunham *et al.*, "Kepler mission design, realized photometric performance, and early science," *The Astrophysical Journal Letters*, vol. 713, no. 2, p. L79, 2010.

[3] D. F. Gray, *The observation and analysis of stellar photospheres*. Cambridge University Press, 2005.

[4] A. Lapedes and R. Farber, "A self-optimizing, nonsymmetrical neural net for content addressable memory and pattern recognition," *Physica D: Nonlinear Phenomena*, vol. 22, no. 1, pp. 247–259, 1986.

[5] J. T. Connor, R. D. Martin, and L. Atlas, "Recurrent neural networks and robust time series prediction," *Neural Networks, IEEE Transactions on*, vol. 5, no. 2, pp. 240–254, 1994.

[6] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[7] ——, "Experimental analysis of the real-time recurrent learning algorithm," *Connection Science*, vol. 1, no. 1, pp. 87–111, 1989.

[8] D. P. Mandic and J. Chambers, *Recurrent neural networks for prediction: learning algorithms, architectures and stability*. John Wiley & Sons, Inc., 2001.

[9] S. S. Haykin, *Neural networks and learning machines*. Englewood Cliffs, Prentice-Hall, New York., 2009, vol. 3.

[10] G. Strang and T. Nguyen, "Wavelets and filter banks," *Wavelets: Theory and Applications: Theory and Applications*, p. 38, 1995.

[11] W. Sweldens, "The lifting scheme: A construction of second generation wavelets," *SIAM Journal on Mathematical Analysis*, vol. 29, no. 2, pp. 511–546, 1998.

[12] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 2008.

[13] C. Napoli, F. Bonanno, and G. Capizzi, "An hybrid neuro-wavelet approach for long-term prediction of solar wind," in *IAU Symposium*, no. 274. Cambridge Univ Press, 2010, pp. 247–249.

[14] ——, "Exploiting solar wind time series correlation with magnetospheric response by using an hybrid neuro-wavelet approach," in *IAU Symposium*, no. 274. Cambridge Univ Press, 2010, pp. 250–252.

[15] G. Capizzi, F. Bonanno, and C. Napoli, "A wavelet based prediction of wind and solar energy for long-term simulation of integrated generation systems," in *Power Electronics Electrical Drives Automation and Motion (SPEEDAM), 2010 International Symposium on*. IEEE, 2010, pp. 586–592.

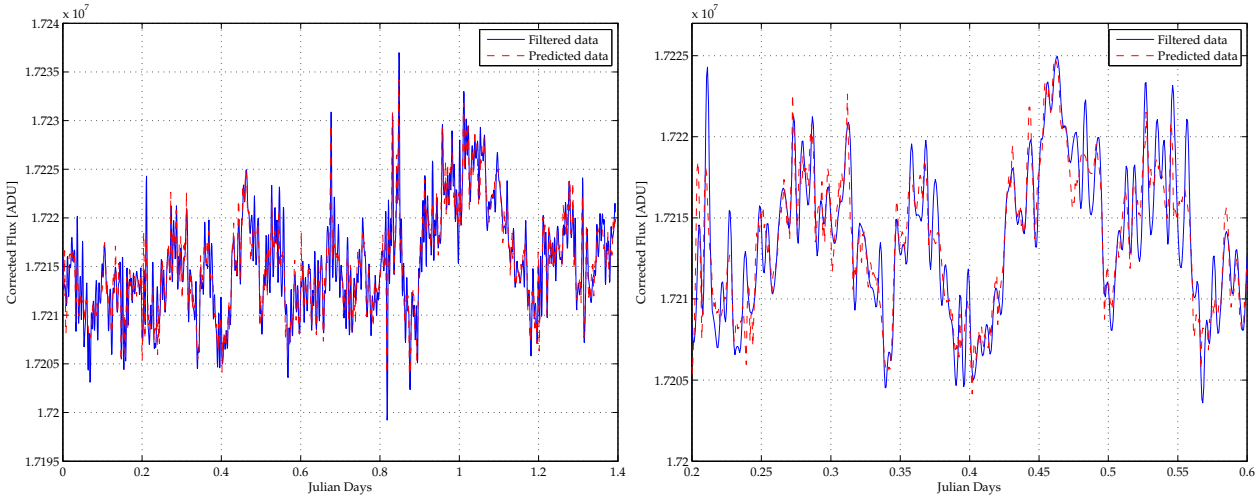

Fig. 5. Simulation results of the forward and backward reconstruction

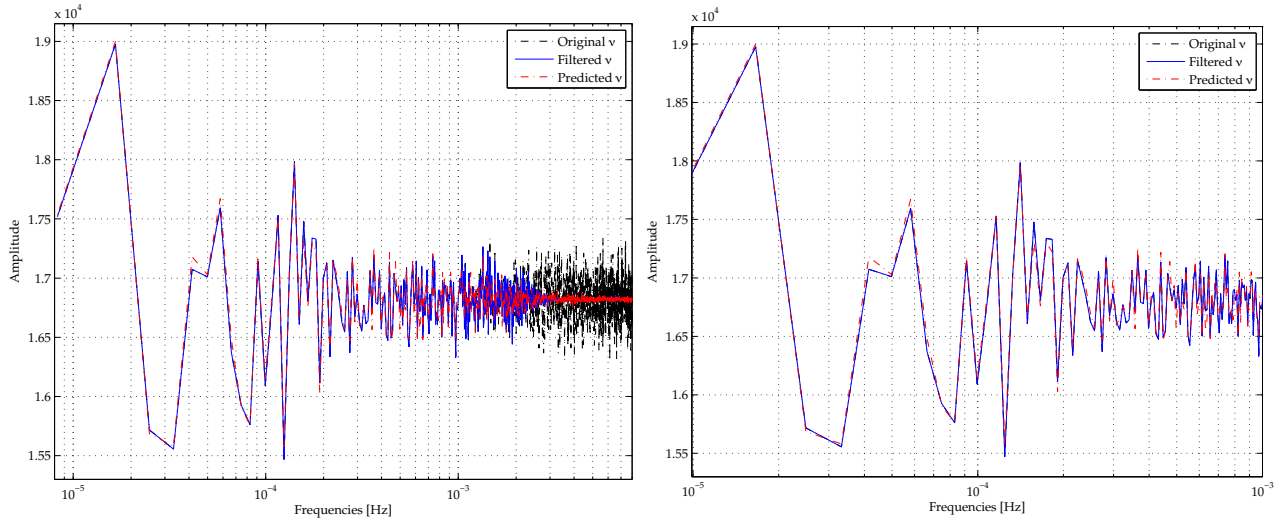

Fig. 6. Simulation results in the frequency domain

[16] G. Capizzi, C. Napoli, and F. Bonanno, "Innovative second-generation wavelets construction with recurrent neural networks for solar radiation forecasting," *IEEE transactions on neural networks and learning systems*, vol. 23, no. 11, pp. 1805–1815, 2012.

[17] M. M. Gupta, L. Jin, and N. Homma, *Static and dynamic neural networks: from fundamentals to advanced theory*. John Wiley & Sons, 2004.

[18] G. Capizzi, C. Napoli, and L. Paternò, "An innovative hybrid neuro-wavelet method for reconstruction of missing data in astronomical photometric surveys," in *Artificial Intelligence and Soft Computing*. Springer, 2012, pp. 21–29.