Prediction of the CATS benchmark exploiting time-reversal symmetry

P. F. Verdes
Institut für Umweltphysik
Universität Heidelberg
Heidelberg, Germany
E-mail: Pablo.Verdes@iup.uni-heidelberg.de

P. M. Granitto, M. I. Széliga, A. Rébola and H. A. Ceccatto
Instituto de Física Rosario
Univ. Nacional de Rosario - CONICET
Rosario, Argentina
E-mail: {granitto,szeliga,rebola,ceccatto}@ifir.edu.ar

***Abstract*— We present a possible strategy for filling the missing data of the CATS benchmark time series prediction competition. Our approach builds upon an appropriate embedding of this time series and the use of bagging of multilayer perceptrons (MLPs). We exploit time-reversal symmetry for prediction within the first four gaps, linking the missing state to symmetrically-located information both in the past and future. One-shot forecasting is then performed for each missing value from distant-enough delays. The suitability of the proposed embedding is assessed empirically by t-testing the goodness-of-fit of models built in symmetric versus asymmetric input spaces. Since this approach cannot be pursued for forecasting the continuation of this time series, in the right end we perform standard, non-iterated forward predictions. Expected error levels are provided according to performance on test data.**

I. Introduction

In this paper we explore a possible forecasting strategy for the CATS benchmark time series prediction competition. Following the usual practice, we start by assuming that the underlying process can be modeled in a d-dimensional pseudo-phase space according to

$$x_{t+1} = f(\boldsymbol{x}_t) + \eta_t,$$

where $\boldsymbol{x}_t = (x_t, x_{t-\tau}, ..., x_{t-(d-1)\tau})$ is the standard homogeneous time-delayed embedding defined by lag τ and dimension d [1], f is the –possibly nonlinear– dynamics to be estimated from the data, and η_t is some residual noise. However, as we shall see in the following Section, after some exploratory data analysis we will build a non-homogeneous embedding and later mirror this framework to exploit time-reversal symmetry. In the last Section we study the performance of ensembles of MLPs in predicting the missing data.

II. Exploratory data analysis

In this Section we will try to answer the following questions: Should we employ nonlinear models? Which predictors could be potentially useful? Finally, we assess the proposed modeling settings by t-testing different null hypothesis.

A. Linearity vs. nonlinearity

Following Casdagli and Weigend [2], [3], we use local linear models to test for nonlinearity. They can be considered as the local Taylor expansion of the unknown f, and are easily determined by minimizing

$$\sigma^2 = \sum_{\boldsymbol{x}_j \in N(\boldsymbol{x}_t,\varepsilon)} (x_{t+1} - \boldsymbol{a}_t \boldsymbol{x}_j - b_t)^2,$$

with respect to $\boldsymbol{a}_t$ and b_t, where $N(\boldsymbol{x}_t, \varepsilon)$ is the ε-neighborhood of $\boldsymbol{x}_t$, excluding $\boldsymbol{x}_t$. This minimization problem can be solved through a set of coupled linear equations, a standard linear algebra problem. Then, the prediction is $\hat{x}_{t+1} = \boldsymbol{a}_t \boldsymbol{x}_t + b_t$. We compute the normalized mean squared error (NMSE) between the original and predicted values, *i.e.*, the MSE divided by the data variance, as a function of the size ε of the neighborhood on which the local linear model is fitted [4]. If the optimum occurs at large neighborhood sizes, then the data will be (at least in the embedding space considered) best described by a global linear model. In contrast, if the optimum occurs at small neighbourhood sizes, then a nonlinear deterministic equation of motion will be a more suitable description of the measured data.

In the lower panel of Fig. 1 we plot the results for the original time series using standard time-delayed embedding vectors of dimensions $d = 2, \ldots 20$. As we can see in this figure, for all the dimensions considered we find the best modeling performances for maximum neighbourhood sizes. This implies, according to the discussion above, that a global linear model best describes the raw data. However, this could be simply due to the high autocorrelation of the big-scale behaviour of the signal. To study the character of the small-scale fluctuations of x_t, we conducted the same analysis on the series of first differences $\Delta x_t = x_t - x_{t-1}$ (shown in the upper panel of Fig. 1). We highlight some interesting features: first, notice that for $d = 2$ (upper curve) the data look like random noise —recall that unpredictability is characterized by a value of NMSE=1. However, in higher dimensions there is some gain in NMSE for small neighbourhood sizes (notice that this series is overall much less predictable than x_t). In conclusion, in this case the results indicate the presence of a weak nonlinear determinism. This suggests that the best strategy might well consist of building first a linear autoregressive model to account for the big-scale behaviour, and then modeling the remaning unexplained variability with a nonlinear predictor (the NMSE levels in the upper panel of Fig. 1 suggest that we may expect to explain up to 20-25% of the remaining variance if we used local linear models). However, we prefer the alternative path of simply gathering together big and small-scale information (x_t and Δx_t, respectively) in a unique state-space description of the system to build a one-shot nonlinear model of x_t.

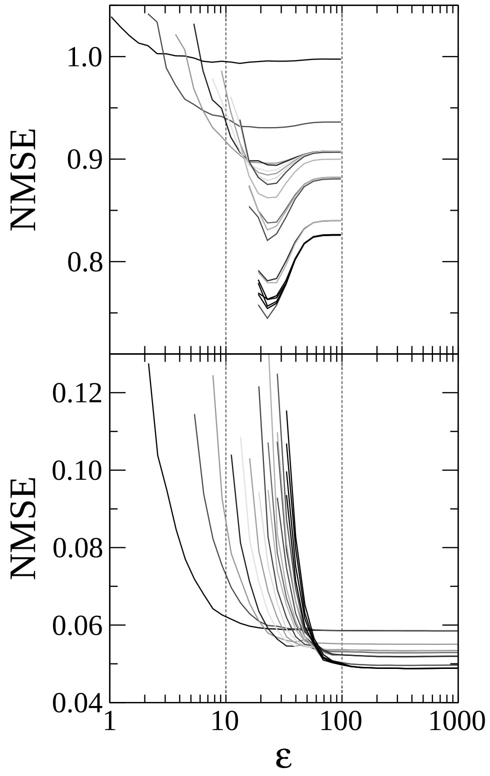

Fig. 1. Normalized mean squared error as a function of neighbourhood size for $\tau = 1$ and embedding dimensions $d = 2$ to 20. Lower panel: original time-series. Upper panel: series of first differences.

B. Selection of relevant predictors

Since the big-scale behaviour seems to be only linearly autocorrelated, for the selection of relevant predictors we focus on the more interesting nonlinear series of first differences. In Fig. 2 we show the decay of the time-delayed mutual information [5] of this series for increasing time-lags τ.

We will consider $\Delta x_{t-\tau}$ to be a relevant predictor of Δx_t if the corresponding MI value exceeds a certain threshold. To fix a concrete lower bound, we rank-order the obtained results for MI (shown in the inset of Fig. 2). We arbitrarily choose the value 0.029, indicated with a dashed horizontal line in both graphs, because after the first 15 most important predictors there seems to be a small step down in relevance —as measured by MI. This procedure pins out $\tau = 1$, 3, 4, 7, 10, 11, 14, 15, 16, 17, 24, 25, 28, 30, and 42.

Although these time-lags were chosen by analysing $\{\Delta x_t\}$, hereafter we will assume that they also provide a sufficiently sampled representation for the more slow, autocorrelated $\{x_t\}$. We thus propose to employ $\{x_{t-\tau}, \Delta x_{t-\tau}\}$ for all τ in the set of 15 lags mentioned above.

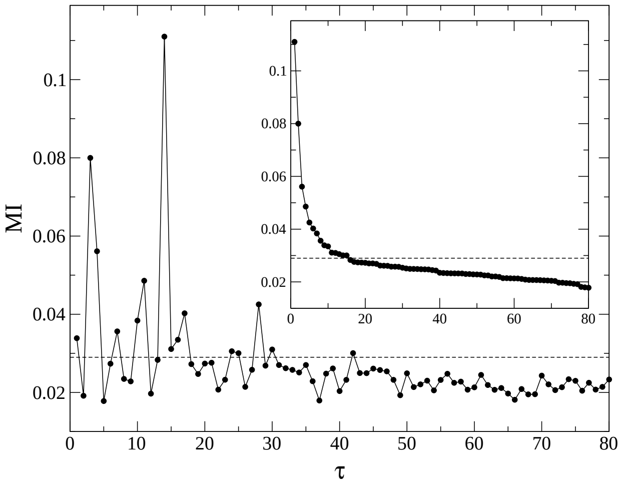

Fig. 2. Time-delayed mutual information (MI) as a function of delay τ for the series of first differences. The dashed horizontal line indicates the threshold employed for selecting predictors.

C. t-Testing the embedding framework

1) Inclusion of $\Delta x_{t-\tau}$: A first check on the proposed strategy was carried out by testing the null hypotheses that the performance of a one-shot nonlinear model obtained from the set of inputs $\{x_{t-\tau}, \Delta x_{t-\tau}\}$ is not better than the one obtained from $\{x_{t-\tau}\}$ only (we employ this condensed notation to indicate that the full set of previously selected 15 time-lags is used). For this we considered single MLPs. We randomly selected 33% of the points in the whole record for testing purposes, and used the remaining 67% data points to train MLPs with architectures $15:10:1$ and $30:10:1$. All the networks were trained until the minimum of the corresponding validation set error was achieved ("early-stopping" criterion). These 2 settings were run on parallel on 100 different random splits of the data, in order to produce a paired t-test.

To study this issue we computed the relative performance differences

$$\Delta E/E = (NMSE_{\{x\}} - NMSE_{\{x,\Delta x\}})/NMSE_{\{x\}}$$

between the test errors obtained using the $\{x_{t-\tau}\}$ set of inputs ($NMSE_{\{x\}}$) and using the full set of inputs ($NMSE_{\{x,\Delta x\}}$). The results of this t-test were 100% conclusive: in all of the 100 experiments the MLPs trained with $\{x_{t-\tau}, \Delta x_{t-\tau}\}$ outperformed the $\{x_{t-\tau}\}$ fitting. In particular, its predictions were, on average, 2.4% better.

2) Inclusion of data from the future: The gap nature of the missing data naturally raises the following question: could a better modeling be triggered by the inclusion of information situated in the future? If so, which "forward" predictors should be employed? To be consistent, we apply the same selection criterion described above. Since MI is by definition symmetrical, *i.e.* $\text{MI}(x,y) = \text{MI}(y,x)$, the time-lagged mutual information of x_t is invariant under time reversal: $\text{MI}(x_t, x_{t-\tau}) = \text{MI}(x_{t-\tau}, x_t)$. Thus the picture for threshold-selection of forward predictors is again Fig. 2, and the same set of 15 relevant delays must be mirrored into the future.

Now we imagine we were to predict a particular missing value within a gap, say, e.g., the point in the 10th position counting from the left border. Since for simplicity we have chosen a one-shot prediction strategy, we restrict ourselves to considering only predictors $\{x_{t-\tau}, \Delta x_{t-\tau}\}$ for $\tau = 10, 11, 14, \ldots 42$ to forecast x_t. More precisely, we ask ourselves: Is this "past" framework better than the one expanded by addition of $\{x_{t+\tau}, \Delta x_{t+\tau}\}$, $\tau = 11, 14, 15, \ldots 42$? To answer this question we have run 100 experiments with different random splits of the data as above, and trained MLPs with architectures $22:10:1$ and $42:10:1$, respectively. The results of the 100 runs are plotted as a histogram in Fig. 3. As can be seen in this figure, this inclusion is responsible for a dramatic difference. In particular, the relative performance improvement is, on average, 50.4%.

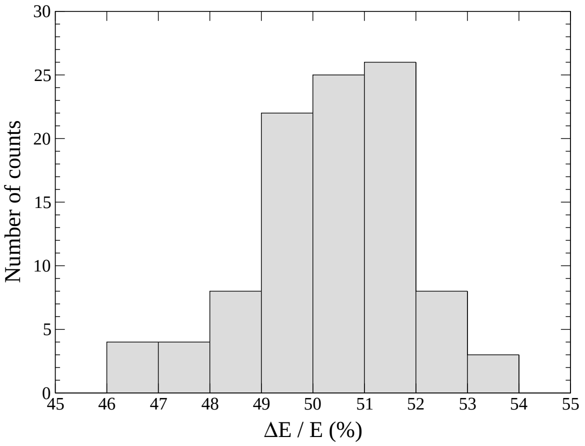

Fig. 3. Relative performance difference $\Delta E/E$ after inclusion of data from the future. The results of 100 independent runs are plotted as a histogram.

III. Filling the Gaps

As follows from the results of the previous Section, it is convenient to employ both time-delayed and time-forwarded coordinates $x_{t\pm\tau}, \Delta x_{t\pm\tau}$. We have individually taylored the concrete embedding setting for each predictee position within the gaps. Numbering them from 1 to 20, in Table I we summarize the temporal location of the predictors involved in each case. Needless to say, for the last 20 points of the whole record we only employed the past delays quoted in the second column of this table.

To produce the concrete final predictions we considered ensembles of MLPs, since they are known to perform better than single networks [6]. Always through t-testing, we compared the advantages of different aggregation strategies [7], [8]. We concluded that bagging [9] was best suited for the problem at hand, showing an improvement over single networks that ranged between 1% and 5%. We found that the error decrease upon aggregation stabilized after a small number of networks, and a size of 10 MLPs was judged to be enough.

The modeling errors obtained in this way over 500 randomly selected test samples are plotted in Fig. 4 as a function of the predictee position within the missing intervals. In the upper panel we depict the expected uncertainties over the first 4 gaps, using predictors both from the past and future. In the lower panel we compare this curve against the expected performance over the 5th gap, which can only be modeled from past data.

TABLE I

Temporal location of the predictors involved in the forecasting of each missing value within the gaps.

Predictee position	Past	Future
1	$-42, -30, \ldots -1$	$24, 25, \ldots 42$
2, 3	$-42, -30, \ldots -3$	$24, 25, \ldots 42$
4	$-42, -30, \ldots -4$	$17, 24, \ldots 42$
5	$-42, -30, \ldots -7$	$16, 17, \ldots 42$
6	$-42, -30, \ldots -7$	$15, 16, \ldots 42$
7	$-42, -30, \ldots -7$	$14, 15, \ldots 42$
8, 9	$-42, -30, \ldots -10$	$14, 15, \ldots 42$
10	$-42, -30, \ldots -10$	$11, 14, \ldots 42$
11	$-42, -30, \ldots -11$	$10, 11, \ldots 42$
12, 13	$-42, -30, \ldots -14$	$10, 11, \ldots 42$
14	$-42, -30, \ldots -14$	$7, 10, \ldots 42$
15	$-42, -30, \ldots -15$	$7, 10, \ldots 42$
16	$-42, -30, \ldots -16$	$7, 10, \ldots 42$
17	$-42, -30, \ldots -17$	$4, 7, \ldots 42$
18, 19	$-42, -30, \ldots -24$	$3, 4, \ldots 42$
20	$-42, -30, \ldots -24$	$1, 3, \ldots 42$

A few comments are in order at this point: first, besides the obvious symmetry, in the upper panel we notice a non-monotonous increase in NMSE as we incursion into the gap from its borders towards its centre. For example, the error in the 7th position is smaller than in the 6th. To understand this behaviour we refer to Table I, where we find that in passing from 6 to 7 we don't loose any past predictors but gain two forward instead, namely x_{t+14} and Δx_{t+14}. In moving one step further on to position 8, the useful information located at $t-7$ is lost and the error must increase again. In the lower panel we observe a monotonous growth in the prediction error on the last gap as the predictee distance to the last known data increases. The stepwise nature of this behaviour is related to the discrete loss of available past predictors (see the second column of Table I). As expected, the results over the first gaps seem to be consistently better than over this last interval.

Finally, in Fig. 5 we illustrate the performance of our completion procedure. To provide a sound basis for a gap filling simulation, we have excluded five equally-spaced block intervals of length 20 from the modeling process described above. We can see in this figure that, as expected, time-symmetric predictions (indicated in black) show a smaller mismatch than one-sided forecasts (grey) near the borders of the gaps. Black and grey curves exemplify possible performances over the first 4 and 5th missing intervals, respectively.

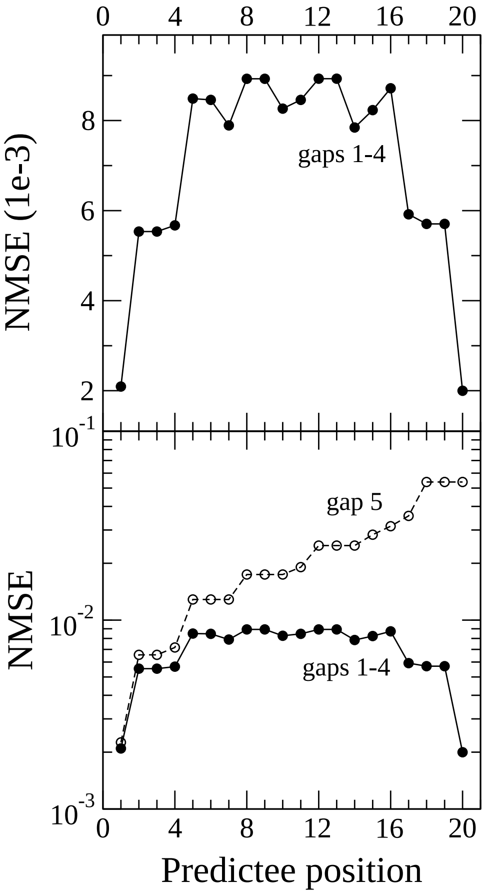

Fig. 4. Estimated distribution profile of modeling errors over the gaps. Upper panel: first four gaps, double-sided embedding. Lower panel, open circles: last gap, one-sided embedding. For comparison, the upper curve is also included (dots).

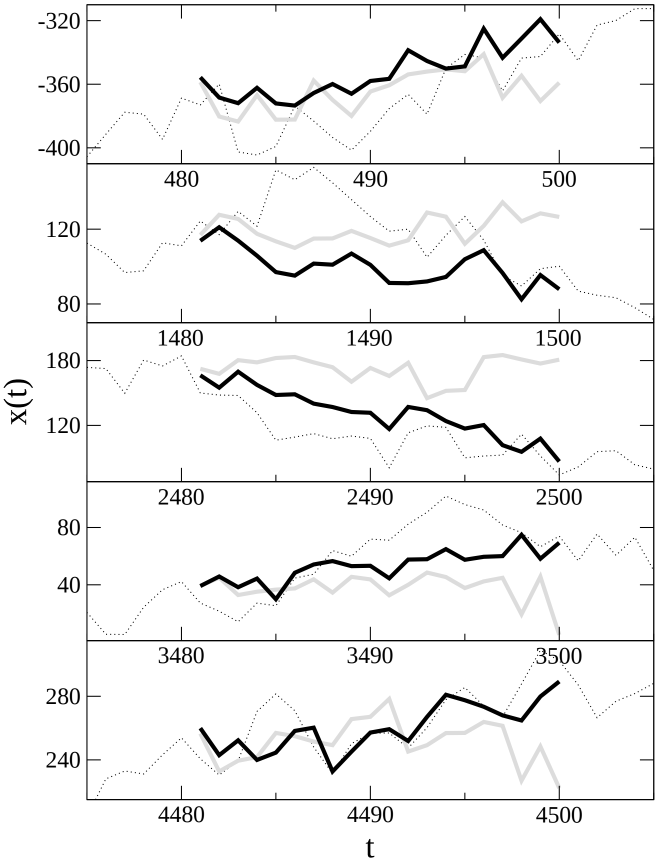

Fig. 5. Examples of imaginary gaps that were completed according to our procedure. True data are indicated by a dotted thin black line, time-symmetric predictions with a full thick black line, and past-based predictions with a full thick grey line.

IV. Conclusions

In this work we have presented a possible forecasting procedure for the CATS benchmark time series prediction competition. Our approach is simple and can be summarized as a time-symmetric embedding of the given time series and a straightforward application of bagging of multilayer perceptrons. One-shot forecasting was performed for each missing value using information both from the past and future, except for the 5th gap where future information is naturally unavailable. Due to time constraints, intuition played an important role in the several choices that had to be made throughout the modeling process. Now, future research should clarify the suitability of these elections. Pending investigations include a more careful selection of possible predictors, a comparison against iterated forecasting, and finally the implementation and comparison against the more involved two-stage modeling strategy suggested by our exploratory data analysis.

Acknowledgments

This work was supported by Fundación Antorchas (P.F.V.) and CONICET (P.M.G., M.I.S., A.R. and H.A.C.).

References

[1] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. Sh. Tsimring, "The analysis of observed chaotic data in physical systems," Reviews of Modern Physics, vol. 65, pp. 1331–1393, 1993.

[2] M. Casdagli, "Chaos and deterministic versus stochastic nonlinear modeling," J. Roy. Stat. Soc. B, vol. 54, pp. 303–328, 1991.

[3] M. Casdagli and A. S. Weigend, "Exploring the continuum between deterministic and stochastic modeling." In Andreas S. Weigend and Neil A. Gershenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, vol. XV of SFI Studies in the Sciences of Complexity, pp. 347–366. Addison-Wesley, Reading, MA, 1993.

[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] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," Phys. Rev. A, vol. 33, pp. 1134–1140, 1986.

[6] A. J. C. Sharkey, editor, *Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems*, Springer-Verlag, London, 1999.

[7] P. M. Granitto, P. F. Verdes, H. D. Navone and H. A. Ceccatto, "A late-stopping method for optimal aggregation of neural networks," International Journal of Neural Systems, vol. 11, pp. 305–310, 2001.

[8] P. M. Granitto, P. F. Verdes and H. A. Ceccatto, "Neural Networks Ensembles: Evaluation of Aggregation algorithms," submitted to Artificial Intelligence, 2003.

[9] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, pp. 123–140, 1996.