TITLE
ASP Conference Series, Vol. **VOLUME**, **PUBLICATION YEAR**
EDITORS

On-line Prediction of Astronomical Seeing Fluctuations with Neural Networks

Alex Aussem & Germain Tran

LIMOS/ISIMA, FRE CNRS 2239 - University Blaise Pascal, Clermont Ferrand II Campus des Cézeaux PO Box 125, 63173 Aubière, France.

Marc Sarazin

European Southern Observatory, Karlschwarzschildstr.2, D85748 Garching,Germany

Abstract. The European Southern Observatory's Astronomical Site Monitor (ASM) for the Very Large Telescope at Cerro Paranal in Chile aims at delivering short-term predictions of the Seeing variability - from 10 up to 60 minutes ahead - to allow optimization of the observing strategy (observing block selection, scheduling of calibration tasks). Years of seeing monitoring at various astronomical sites have shown that the seeing is not stationary and motivated the definition of a variability measure, the FEFSC, i.e., the "Finite Exposure Fractional Seeing Change" (Racine 1996, Sarazin 1997), the impact of which on the performances of adaptive optics systems has been clearly illustrated (Rigaut & Sarazin 1999). Extensive data collected at Paranal have been used to appraise a forecasting methodology based on neural networks (NN). We show that NN achieve a success rate up to 70 % in predicting the FEFSC trend. The forecasting tools which have been prototyped, are a contribution to a future telescope decision support system.

1. Introduction

In this article we discuss on-going work carried out in the area of seeing variability prediction at Paranal, the new site of the European Southern Observatory (ESO) in Chile. The seeing, defined as the 'full width at half maximum' of the image at 500 nm at zenith, is a measure of the atmospheric diffraction of the light waves. Of great importance for the astronomers is the seeing fluctuations for instance when calibration tasks are performed some time before or after the observation. Unfortunately, years of seeing monitoring in various astronomical sites show that the seeing is not stationary, but that its average value fluctuates (Racine 1996; Muñoz-Tuñón, Vernin & Varela 1997; Sarazin 1997). In turn, these fluctuations can induce errors in science observations, see (Rigaut & Sarazin 1999) for a discussion of the impact of seeing changes on the performances of the deconvolution of adaptive optics images. Therefore the need for a prediction engine capable of providing short-term forecasts of the seeing fluctuations.

Since 1998, a monitoring station has systematically been tracking a range of meteorological and atmospheric variables at Paranal. The continuous expansion of the automatic data acquisition process affords an opportunity to apply nonlinear regression techniques such as NN. The practical experience and theoretical knowledge of the meteorologists have limitations, which points to the inherent difficulty of deriving rules for local seeing forecasting. Therefore, prediction is sought without reference to any underlying physical model. The meteorological variables are assumed to carry within them the potential for predicting some information on the seeing fluctuation over short-term periods.

The use of nonlinear models is strongly motivated by the work carried out by (Murtagh & Sarazin 1993; Aussem, Murtagh & Sarazin 1994; Aussem, Murtagh & Sarazin 1995) which shows that NN methods provide encouraging results in terms of seeing and local temperature forecasts. In this paper, we examine the effectiveness of neural network to infer the rules that govern the short-term seeing variability over few minutes from the current and past meteorological measurements. We first discuss the prediction models we have employed. We then address the seeing requirements in astronomical observations and we examine the issues we faced in building a consistent data set for 10 up to 60 mn ahead predictions.

The neural-based prediction software should ultimately deliver in real time, and in an automated manner, the variability of the seeing in the next 10, 20, 30, 40, 50 and 60 mn. The system is to be integrated into the existing package of the ASM forecasting tools,

2. Seeing requirements in astronomical observing

Service observing is an increasingly popular mode of observation which does not require the presence of the end user and thus allow an increased flexibility in the scheduling of the observing blocks. The observing queue is organized by fitting at best the initial user requirements to the current observing conditions. A prediction of the seeing variability other the next hour would allow the queue to be optimally managed: short observations are preferable when the seeing is highly variable and inversely.

The seeing is defined as the Full Width at Half Maximum of the image at 500 nm, at zenith. We call this quantity FWHM. Seeing data is collected at Paranal since 1989, with a Differential Image Motion Monitor (DIMM). Except for high winds, it is bias free and provides an rather accurate measurement of the seeing (± 10 %). In the following we use data collected between September 1st 1999 and September 30th 1999. Meteorological data is collected on a minute basis all day whereas the seeing is collected on a minute basis during nighttime only from about UT 10 PM to 10 AM. A log normal distribution was shown to provide a satisfactory fit to the seeing histogram (Rigaut & Sarazin 1999).

The seeing displays variability at all time scales, up to several hours. There are rapid and sometimes large variations of seeing on time scales of minutes. Astronomers are mostly interested in these short-term variations. One way to characterize these temporal seeing changes is the Fractional Seeing Change (FSC) introduced in (Racine 1996). For our operational purpose, we define a

Finite Exposure Fractional Seeing Change (FEFSC)

$$\mathrm{FEFSC}(t,\tau) = 2\frac{\sum_{i=0}^{\tau}\mathrm{FWHM}(t+\Delta t-i) - \mathrm{FWHM}(t-i)}{\sum_{i=0}^{\tau}\mathrm{FWHM}(t+\Delta t-i) + \mathrm{FWHM}(t-i)} \tag{1}$$

The FEFSC is a convenient quantity, easily understandable, which expresses the relative seeing change on a given time scale. Zero is the ideal value for the FEFSC . $FWHM(t)$ is the seeing value at time t. Instantaneous seeing values are meaningless since any science exposure takes a laps of time, therefore we consider average seeing values over a time period τ equal to 10 mn here (i.e. over 10 seeing values at a rate of one per minute).

3. Neural networks

It is difficult to present the connectionist network, especially when the audience consists of a mixture of meteorologists and astronomers some of whom will know a great deal about connectionist approaches to modeling and prediction, while the knowledge of others is limited in scope. However, we have tried to highlight the specificities in a brief review in which neural networks are analyzed as regression models.

A multi-layer neural network (also called Perceptron or feed-forward network) is a simple, well- known and in-depth studied input-output (i.e., static) model, with non-linear transfer functions, offering universal function approximation capability. Such a model can accommodate a combination of continuous (usually interval-scaled) and discrete numeric variables even if, in practice, they have severely heterogeneous certainties. Additionally, classical neural paradigms are global models, that is, a single neural model is trained to model the entire simulation response surface. This differs from polynomial regression for instance, where the regression surface is fitted locally. Finally, the parallel architecture provides robustness to incomplete or erroneous data sets and offers fault tolerant, real-time performance.

It is important to stress that the optimal mean-squared predictor is nothing else than the conditional mean of the desired output, given the input. This intuitive solution is also the closed-form optimal solution for a neural network trained using least squares. It should be understood, however, that a neural network of fixed size can only approximate the optimal function thus introducing an effective bias in the solution. Furthermore, training on limited data results in an effective variance over the possible converged solutions, we say that the model has overfitted the data. A larger network has smaller bias but requires more samples to train and so has an effectively larger variance. This trade-off is typical in regression theory.

What makes neural networks better at finding solutions than splines for instance is a far more difficult problem to answer and requires thorough statistical understanding of the notions of consistency and generalization. But it is well beyond the scope and purpose of this paper to delve deeper into statistical theory of optimal predictors. Readers interested in more theoretical material are encouraged to consult the literature. In the next section, we discuss a more general neural architecture suited to our forecasting problem.

3.1. Dynamical Recurrent Neural networks

In this study, we experiment with so-called Dynamical Recurrent Neural Networks (DRNN) (Aussem, Murtagh & Sarazin 1995, Aussem 1999), an extension of the traditional Perceptron endowed with some internal memory in order to enable the model to capture some information from the past. DRNN are a class of fully recurrent neural networks obtained by modeling synapses as autoregressive filters, and therefore, involves explicit inclusion of delays in the system equations. By virtue of their internal dynamic, DRNN approximate the underlying law governing the time series by a system of non-linear equations of internal variables. They therefore provide history-sensitive - or context sensitive - forecasts without having be explicitly fed with external memory. Therefore, the input/output mapping becomes dynamic. The implementation of the DRNN is that published in (Aussem, Murtagh & Sarazin 1995, Aussem 1999) which is tailored to time series prediction.

3.2. Training set-up

Before we turn to the experiments, we detour to review two (very basic) techniques we found to be useful for networks to run optimally. In an attempt to optimize the learning rate η automatically, we allow the it to vary in the course of training since the best values at the outset of training may not be so good later on. Our approach is to check whether the weight updates at each iteration actually decrease the cost function. If not then the process is overshooting and η should be reduced. On the other hand, if several steps in a row have actually decreased the cost function, then η is increased. It appears best in the literature to increase η by a constant and to decrease it geometrically to allow rapid decay when necessary. This gives the overall scheme

$$\eta = +\alpha \text{ if } \Delta \mathrm{E} \leq 0 \tag{2}$$

$$\eta = -\beta\eta \text{ if } \Delta \mathrm{E} > 0 \tag{3}$$

where α and β are appropriate time constants. In this paper, we took $\alpha = 10^{-3}$ and $\beta = 0.9$.

In order to prevent data overfitting (i.e., a small bias at the expense of a large variance), a topological reduction approach was adopted in these studies. The approach consists in pruning appropriately non-useful connections during training. This is achieved by giving each connection a tendency to decay to zero, such that

$$\omega_{ij}^{new} = (1 - \epsilon)\omega_{ij}^{old} \tag{4}$$

so that the connections disappear unless reinforced (Weigend, Rumelhart & Huberman 1990). For small ω_{ij}'s to decay more rapidly than larger ones, we made ϵ dependent on ω_{ij} by

$$\epsilon_{ij} = \frac{\gamma\eta}{(1 + \omega_{ij}^2)^2} \tag{5}$$

where γ was varied in the course of training. We found out that to it is useful to begin with $\gamma = 10^{-3}$ until the error reaches a minimum and then to decrease the value down to 10^{-4} so the error continues to decrease at a slow rate. We just stress that more sophisticated and powerful topological reduction approaches exist. They consist in appropriately pruning (or penalizing) non-useful connections as training proceeds, preventing the network from overfitting the data. For further details, the reader is directed to the references.

4. Performance criteria

To gauge the performance of the forecasting, we use 3 distinct and complementary criteria. The Mean Squared Error (MSE) over N examples is defined below where v_i is the prediction and t_i is the target.

$$MSE = \frac{1}{N}\sum_{j=1}^{N}(v_j - t_j)^2 \tag{6}$$

The hit rate measures our ability to anticipate the right trend. In other words, the hit rate is the probability that the prediction and the target have the same sign.

Table 1. Contingency table. The hit rate is defined p11 + p22, i.e., when the prediction and the target have the same sign. High means positive FEFSC and Low means negative FEFSC.

Forecast/target	High	Low	Sum
High	P_{11}	P_{12}	$P_{11} + P_{12}$
Low	P_{21}	P_{22}	$P_{21} + P_{22}$
Sum	$P_{11} + P_{21}$	$P_{12} + P_{22}$	

The cross-entropy between X, the prediction, and Y the target, denoted by $I(X,Y)$, is is a well known information measure. Let $H(X)$ denote the Shannon entropy of the random variable X. The cross entropy is employed here to quantify the amount of information the prediction provides on the true value. It can be calculated from the contingency table by,

$$\begin{aligned}
H(X) &= -(p_{11}+p_{21})\log p_{11}+p_{21} - (p_{12}+p_{22})\log p_{12}+p_{22} && (7)\\
H(Y) &= -(p_{11}+p_{12})\log p_{11}+p_{12} - (p_{21}+p_{22})\log p_{21}+p_{22} && (8)\\
H(X,Y) &= -p_{11}\log p_{11} - p_{12}\log p_{12} - p_{21}\log p_{21} - p_{22}\log p_{22} && (9)\\
I(X,Y) &= H(X)+H(X/Y) = H(X)+H(Y)-H(X,Y) && (10)
\end{aligned}$$

5. Data selection

In this section we mainly examine the problems raised by the construction of consistent training sets for seeing variability prediction at 10 to 60 mn intervals. Provided the meteorological conditions are good, the seeing variability which we have to anticipate is the major indicator of the quality of observing conditions.

The meteorological data are archived at about the same frequency than that of the seeing but the latter is only recorded at night. The meteorological variables are monitored every 2 seconds and then aggregated to form the average, standard deviation, minimum and maximum, over 60 seconds and 20 mn intervals. Data are archived in real time 24 hours a day. We will mostly rely on average measurements computed over 60 seconds.

Based on expert knowledge, we have ranked the variables according to their expected importance given our prediction task. Among the most important variables are: TA1S (air temperature), RH1S (humidity), WD1S (wind speed), WS1S (wind direction), PA1S (pressure). Subscript *1* means *highest sensor level (30 m)*, and *S* means *average over 60 s*. Then come the vertical gradients, represented by TA1S-TA2S (Tair30-Tair2m), TA1S-TG1S (Tair-Tsol), RH1S-RH2S, WD1S-WD2S, and WS1S-WS2S. The standard variations over 60s are also available; they express a situation change, especially for the variables: TA1S, RH1S, WD1S, WS1S, PA1S. Finally, the vertical wind speed, UW3S, might be of interest. Finally, Sc is the stellar scintillation which is viewed as an indicator of turbulence at high altitude. Other parameters are available but were not considered useful in regards to the seeing fluctuation.

A Principal Component Analysis (PCA) (Stauffer & Moerel 2000) was first carried out to find out the most relevant variables. 8 variables were selected among the above variables: TA1S, RH1S, TA2S- TA1S, WS1S, sin(WD1S), WS2S-WS1S, UWS3 and FWHM. Among these 8 variables, extensive simulations were conducted using 11 different combinations of 3 up to 4 variables for the input vector to a neural network of fixed size. It is worth remarking that the more irrelevant the inputs, the more difficult is the training. In addition, it should be clear that the computational time is a growing function of the number of inputs and delays linking these inputs to the hidden neurons. Therefore, we had interest to keep the number of inputs as small as possible.

Extensive comparisons were made in terms of MSE on training and test sets, hit rate and cross entropy between the target and the 10 mn ahead prediction in order to empirically select the set of input variables exhibiting the best performance. Coding the data plays a pivotal role, the inputs were centered and normalized. For each setting (i.e., a set of input variables), we ran 10 trials with different seeds for the initial weights and different training sets of size 3000 and test sets of size 500 spanning the 10 first days of September 1999, see (Tran 2000) for details. The average, the minimum and maximum MSE, hit rate and cross entropy were calculated. The hit rate varied between 63 % and 67 %, the MSE on test set between 0.0131 and 0.0147, and the cross entropy between 0.038 and 0.062. In view of these comparative results, we selected 4 variables, among the 8, that yielded the best results:
-TA1S: temperature at 30 meters above ground,
-TA1S-TA2S: difference between temperature at 2 meters and 30 meters above ground,
-WS1S: wind speed at 30 meters above ground,
-FWHM: seeing value.

6. On-line training

As we are seeking to achieve on-line prediction of the FEFSC, we have to devise an automated system that fills in the DRNN training set with newly recorded data, and meanwhile, remove the oldest data. Our goal is to accomplish fast training while preserving good performance. Provided some care is taken over the way in which the training set is updated, the DRNN models is allowed to adapt to contextual changes while retaining useful information from the past.

On the one hand, the training set should be large enough compared to the number of parameters to alleviate the problem of overfitting. On the other hand, it should not be too large for the training to be exceedingly slow. In addition, there are obvious non stationarities in the data, owing to the seasonal effects that have to be taken into account. Therefore, we have split the data in groups of 500 successive patterns, representing more or less a daily record. The performance were gauged with different sizes of training sets. We found a training set of 3000 patterns (i.e., about 6 days) to be optimal in terms of MSE performance over a few days. The training set was shifted everyday by 500 values. That is, the oldest 500 values among the 3000 were replaced by the newer 500 ones. With this strategy, the neural networks is re-trained once a day. The proposed method loosens the computational burden by orders of magnitude while preserving good performances.

7. Experiments and validation

Extensive simulations were conducted to assess the performance of the neural-based on-line prediction scheme in an operational setting. Selection of candidate DRNN architecture was based on trial and error along with various heuristics. Insufficient time and resources were the major stumbling blocks. Broadly speaking, the computational complexity growths with the squared number of synaptic weights. Therefore, we were forced to limit the number of neurons, delays and inputs as discussed before.

The selected DRNN architectures for 10 up to 60 mn are all made up from 4 inputs, 6 hidden units and a single output unit providing the FEFSC prediction. Inputs are fully connected to the hidden units; the latter are in turn fully connected to the output unit. The maximum delays from the inputs to the hidden units vary from 30 to 60 mn. Delayed feedback connections from the output to the hidden units were used to back-propagate the oldest available targets back into the network. The DRNN models have about 650 parameters to be adjusted, i.e. the weights. The delays are fixed. Note that we feed back the targets instead of the predictions. This is called *teacher forcing* learning. The neural system can be thought of as a non-linear extension of an auto- regressive moving average (ARMA) process using delayed inputs and delayed targets as the current input to predict the next target.

We have analyzed the 10 mn up to 60 mn ahead forecasts of the FEFSC over the test sets during September 1st, 1999. The performance results are condensed in Table 2. Note that the trivial carbon copy predictor achieves no more than 45 % hit rate on any time horizon (worse than a random guess), and the cross entropy is no more than 0.005. Therefore, the neural based engine provides encouraging results. The prediction did not degrade much from the training to the test sets, indicating good generalization ability. It is also worth remarking that the hit rate is maximal for 60 mn prediction horizon, and minimal for the 20 mn horizon. Besides, the MSE is not proportional to the hit rate: the highest MSE on the test set is obtained with the DRNN that achieves the highest hit rate.

Table 2. The performance results over one night of the 10 to 60 mn networks trained on their respective 6 nights sets.

	10 mn	20 mn	30 mn	40 mn	50 mn	60 mn
MSE train	0.014	0.032	0.043	0.047	0.055	0.049
MSE test	0.014	0.034	0.039	0.057	0.057	0.085
Hit rate (%)	67	62	64	68	69	71
I(X,Y)	0.062	0.032	0.043	0.061	0.073	0.071

8. Conclusion

Extensive data collected at Paranal have been used to appraise an on-line forecasting methodology based on (dynamical recurrent) neural networks. We obtained a success rate up to 70 % in predicting the FEFSC trend, i.e., the *Finite Exposure Fractional Seeing Change*, from 10 up to 60 mn ahead, on the basis of 10 mn exposures. The forecasting tools which have been prototyped, are a contribution to a future telescope decision support system.

References

Aussem, A., Murtagh, F., & Sarazin, M. 1994 in Vistas in Astr. **38**, 357-374.

Aussem, A., Murtagh, F., & Sarazin, M. 1995 in Int. Journal on Neural Systems 6, 145-170.

Aussem, A., Murtagh, F., & Sarazin, M. 1996 in Neurocomputing 12, 359-373.

Aussem, A. 1999 in Neurocomputing 28, 207-232.

Muñoz-Tuñón, C., Vernin, J., & Varela, A 1997 in A&AS**125** 183-193

Murtagh, F., & Sarazin, M. 1993 in PASP 105, 932-939.

Murtagh, F., Aussem, A., & Sarazin, M. 1995 in PASP **107**, 702-707.

Stauffer, G., & Moerel, F. 2000 in Master of Science Report (in french), Isima/University Blaise Pascal, Clermont- Ferrand, France.

Racine, R. 1996 in PASP 108, 372-374.

Rigaut, F., & Sarazin, M. 1999 in ESO/OSA Topical Meeting on Astronomy with Adaptive Optics, Sonthofen, Germany

Tran, G. 2000 in Master of Science Report (in french), Isima/University Blaise Pascal, Clermont-Ferrand, France

Weigend, A.S., Rumelhart, D.E., & Huberman, B.A. 1990 in International Journal of Neural Systems **1**, 195-220.