Investigation of the Optimum Minimum Input Data For the Forecasting of 3D Point Position Changing, Using Non-Linear Autoregressive Neural Networks

Alevizakou Eleni-Georgia[1,*], Evangelia Lambrou[2]

[1] School of Rural and Surveying Engineering, National Technical University of Athens, 15780 Zografos, Athens, Greece (elalevizakou@gmail.com)
[2] School of Rural and Surveying Engineering, National Technical University of Athens, 15780 Zografos, Athens, Greece (litsal@survey.ntua.gr)
[*] Correspondence: elalevizakou@gmail.com

Key words: *Artificial Neural Networks (ANNs); Geodesy; position changing; forecasting; non-linear autoregressive recurrent network (NAR); non-linear autoregressive with eXogenous inputs (NARX)*

ABSTRACT

One of the most attractive and popular intelligent technique in the scientific community, lately, is the Artificial Neural Networks (ANNs). They have been globally used in most disciplines (e.g., Economics, Medicine, Engineering) in order to solve difficult and complex problems neither as a new method nor as a complementary, more effective one. As it was expected, ANNs have been also introduced in numerous geodetic fields such as regional mapping of the Geoid, sea level forecasting, coordinate transformation, deformation monitoring etc. As far as the deformation monitoring is concerned ANNs have recently been used in the position changing forecast problem. In particular, a methodology, based on Non-Linear Autoregressive (NAR) and non-linear autoregressive with eXogenous inputs (NARX) Neural Networks, has been developed in order to provide forecasts for point position changing. A question that emerged applying this methodology was the investigation of the optimum minimum number of available input data. The research reported in this paper aims to investigate this number, so that the produced short-term and long-term forecasts would have acceptable Mean Absolute Error (MAE), depending on the order of the individual position changing. The idea behind this research is mainly based on the probability of lack of access to big data. Experiments with different number of daily continuous data from GNSS permanent stations, starting from 3190 daily records, were carried out. Results reveal that the number of inputs plays an important but not crucial role during implementation of the methodology. It was concluded that reliable forecasts could indeed be produced using smaller number of input data.

I. Introduction

For decades, the forecasting issue in general has been a very attractive and active topic of different research fields. As a result, series of studies related to these problems, were conducted. Some of the most popular scientific fields, in which the researchers mainly focus on, are the following: Economics (Moshiri S. et al., 2000; Demir et al., 2015), Meteorology (Abhishek K., 2012), Environmental Area (Babu B.V. et al., 2003; Khalil B.M. et al. 2012) etc. In these fields due to the difficulty and complexity of each examined problem, it is often considerably challenging to find the proper forecasting technique.

In order to generate the needed forecasts, with the desired and acceptable accuracy, many methods and techniques have been developed, including intelligent ones (Artificial Intelligence, AI). Lately, Artificial Neural Networks (ANNs) constitute one of the most attractive technique due to the excellent results and noteworthy success achieved in a number of applications and researches.

One complicated problem in which ANNs are recently used is the forecasting of 3D point position changing. According to a published methodology (Alevizakou E. et al., 2018), non-linear autoregressive recurrent networks (NAR) can be used for forecasting one time step 3D point position changing, and non-linear autoregressive with eXogenous inputs (NARX) for multi-step forecasting. Briefly, this methodology follows the following basic steps:

1. Problem definition
2. Finding the available data in time series form
3. Preliminary data analysis (data pre-processing)
4. ANN definition
5. Evaluation of forecasts.

However, by applying this methodology a question that emerged was the investigation of the optimum minimum number of available input data. The idea behind this research was the fact that most of the forecasting methodologies, which use ANNs, are based on the existence of a large number of available historical data. Unfortunately, a fact that creates problems to the researchers is the possibility of access to big data means long-term continuous daily GNSS coordinates (time series), which is not always an option. That lack of data was the main idea behind this research. Specifically, this work focuses on the investigation of this number when forecasting of 3D point position changing is concerned.

II. THE ANN-BASED METHODOLOGY

A. *Research methodology*

The research methodology, in which this work is based on, can briefly shown in the following flow chart.

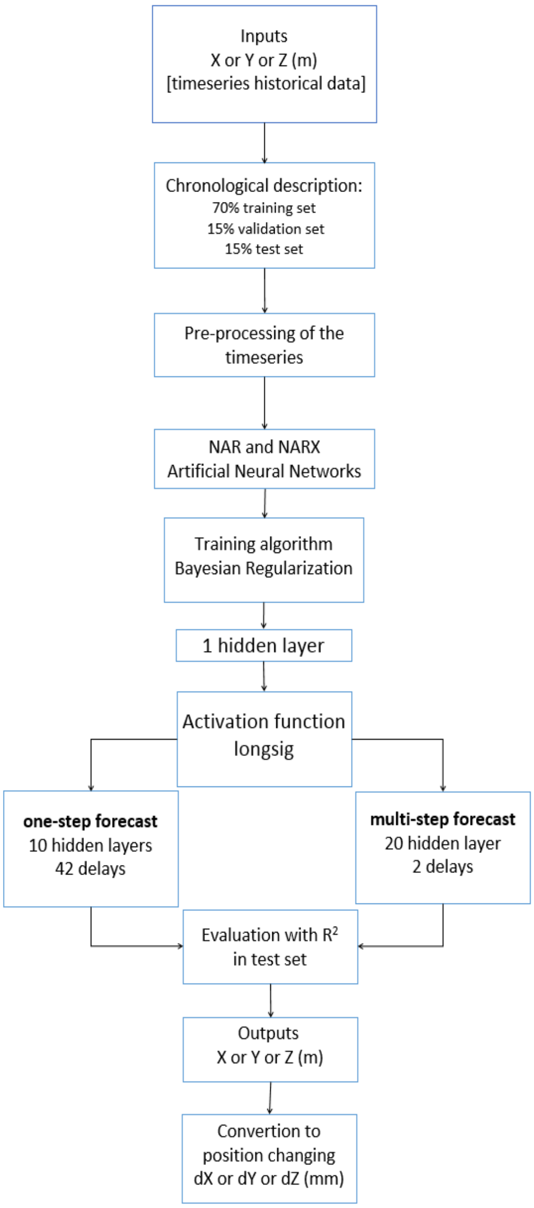

Figure 1. Flow chart of the ANN-based methodology

The main concept is the use of the non-linear autoregressive recurrent networks (NAR) for forecasting one time step and non-linear autoregressive with eXogenous inputs (NARX) for multi-step forecasting (Allende H. et al., 2002; Cai S. et al. 2012; Mounce S.R., 2013 Safavieh E. et al., 2007).

In the case of NAR, the forecasting of the $\ell(\mathrm{t})$ element of the time series (in this study initially the X, Y or Z coordinate and subsequently the desired position changing dX, dY or dZ) is made, using only "d" past values of that time series, according to the equation:

$$\ell(\mathrm{t}) = \mathrm{f}(\ell(\mathrm{t}-1),\ldots,\ell(\mathrm{t}-\mathrm{d})) \tag{1}$$

Similarly, in the case of NARX, the "d" past values of the same time series "ℓ" and "d"of another exogenous time series "q" are used, according to the equation:

$$\ell(\mathrm{t}) = \mathrm{f}(\ell(\mathrm{t}-1),\ldots,\ell(\mathrm{t}-\mathrm{d}),\, q(\mathrm{t}-1),\ldots,q(\mathrm{t}-\mathrm{d})) \tag{2}$$

The proposed ANN consists of one hidden layer and has as activation function of hidden neurons and output neurons the logsig function. Data is separated in chronological order (70%, 15%, and 15%) and their normalization is in the space [0.1, 0.9]. The training algorithm is Bayesian regularization. The number of hidden neurons is 10 and the number of delays 42 for one-step forecasts. While for multi-step forecasts, there are 20 neurons and 2 delays. In addition, X, Y and Z coordinates have been decided to be separately the inputs, with the corresponding outputs.

When the above methodology is applied with the proposed ANN, it has been proven that the result is the production of reliable forecasts of 3D point position changing data.

The reliability of every forecast is examined by the selected evaluation criteria. There is a great number of possible criteria, which can be used (Schroeder et al. 2009; Erdogan 2010; Yilmaz et al., 2014; Alevizakou et al., 2017).

The two proposed criteria are the Mean Absolute Error (MAE) and the Coefficient of Determination (R^2), which can be calculated using the following equations:

$$(\mathrm{MAE}) = \frac{1}{\mathrm{n}} \cdot \sum_{\mathrm{t}=1}^{\mathrm{n}} |\mathrm{e_t}| \tag{3}$$

$$(\mathrm{R}^2) = \frac{\mathrm{RSS}}{\mathrm{TSS}} = \frac{\mathrm{TSS\text{-}ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{ESS}}{\mathrm{TSS}} \tag{4}$$

where $\mathrm{e} = \mathrm{D\text{-}F}$ the error between the "n" original data (D) and the forecasted (F).

$\mathrm{TSS} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{D_i} - \bar{\mathrm{D}})^2$ the total sum of squares

$\mathrm{ESS} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{F_i} - \bar{\mathrm{D}})^2$ the regression sum of squares or Explained sum of squares

$\text{RSS}=\sum_{i=1}^{n}(D_i-F_i)^2$ the sum of squares of residuals or Residuals sum of squares

Specifically, the closer to one (1) the R^2 is, the better the forecast. A forecast can also be considered as accurate if the MAE is significantly smaller than the mean forecasted position change. MAE is proved a satisfied and unbiased index as assigns equal weight to all errors when calculating overall performance, so it provides a real estimation of the forecast uncertainty.

III. Case study

In this research, daily records from permanent GNSS (Global Navigation Satellite System) stations from the Plate Boundary Observatory (PBO), originally from the year 2013 to 2018, are used. It was decided to use this kind of data because scientists use them in numerous of studies for the monitoring of points position in constructions or for the Earth's crustal movements (Bell J.W. et al., 2002; Chan W.S. et al., 2005; Psimoulis et al., 2007; Farmer G.T. et al., 2007). At the same time, these data are available free online (http://www.earthscope.org/node/395; http://www.unavco.org/projects/majorprojects/pbo/pbo.html) which is considered as one of their advantages. The other advantage of these data is that they can be considered as "big data" because they consist of large quantity volume. However, this study examines cases where these data are not available. The main idea was the inability to access large volume data, like the data mentioned above.

A. *Data*

In this particular study, data from three (3) GNNS stations was used in order to have trusted conclusions. Initially, the PBO permanent GNSS network is composed of more than 1000 stations, but when so many data are available, it is crucial to find a way to get a sample to make the experiments.

In the international literature, there are three methods to obtain a representative sample (Elsayir H.A., 2004; Kareem A. et al., 2015) and by following them, three kinds of samples are created: simple random sample, systematic or quasi-random sample and stratified random sample. In this research, it was decided to use a stratified random sample and the variance of them to be used as the criterion. The variance is calculated from the next formula:

$$s_\ell^2=\frac{1}{n-1}\cdot\sum_{t=1}^{n}(\ell_t-\bar{\ell})^2 \quad (5)$$

Where n = the number of the original data

ℓ_t = the element of the time series (the X, Y or Z coordinate).

The variances of the X, Y and Z time series are divided into three (3) categories in order to rank the GNNS stations into three groups. The first category is the one with the smallest variances from 0 to 1mm, the second with medium variances from 1.1mm to 5mm and the last one with the large variances from 5.1mm to 10mm. From each category, a GNSS station is randomly selected.

After this procedure the final sample, that is used, is represented in the following table (Table 1).

Table 1. Variances of the selected stratified random sample

Station name	***Variance X (mm)***	***Variance Y (mm)***	***Variance Z (mm)***
P788	0.3	0.3	0.2
AC57	1.1	3.0	1.1
P515	6.0	6.0	3.0

The time series of the AC57, P515 and P788 stations started from 23.06.2006, 10.05.2006 and 29.08.2008, respectively. The main disadvantage of these time series is that they included some "problems". The main problems are the followings : the missing values presented in the time series, the possible zero values in the event that it is known that they should not exist, the duplicate recordings that may exist and any unusual observations (outliers) (Alevizakou E.G. et al., 2017). These problems must be managed during the pre-processing stage of the methodology. After the solution of the above-mentioned problems the remaining number of daily data from each GNSS permanent station, is 3190 records.

B. *Empirical findings/ Results*

As it is already mentioned, the main idea for this study, is based on the probability of lack of access in big data, such as the GNSS daily records from a free access permanent stations' network. Experiments with different number of daily data, starting from 3190 records, were carried out. Nevertheless, it was decided that 3190 daily records (that is almost 8 years) is a number that is not easy to achieve, so the experiments were started with a smaller number of 730 daily records (3 years).

The experiments vary in the number of inputs, starting from the 730 newest chronologically data (day 2460 to day 3190) and finishing to the 100 newest daily data (day 3090 to day 3190). The results are short-term and long-term forecasts and the criteria, which is used, is the R^2 of the training set. The first three figures (figure 1, 2 and 3) show the R^2 for short-term forecast of the X, Y and Z coordinate respectively. In addition, the next three (figure 4, 5 and 6) for multi-step forecasts.

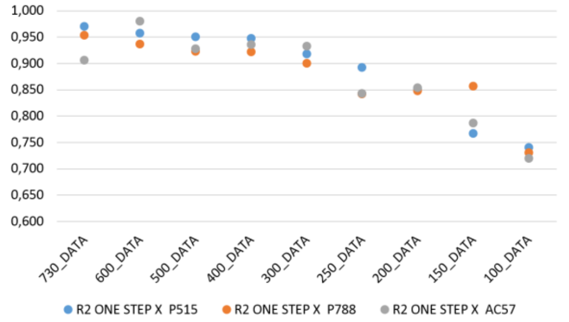

Figure 1. Results of R^2 for the short-term forecast of the X (m) coordinate.

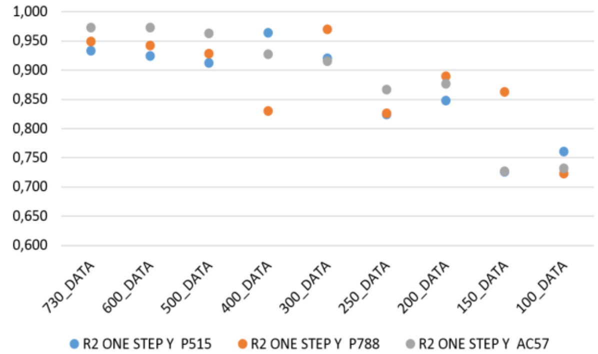

Figure 2. Results of R^2 for the short-term forecast of the Y (m) coordinate.

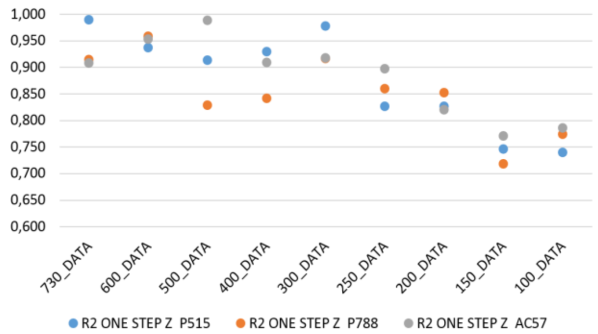

Figure 3. Results of R^2 for the short-term forecast of the Z (m) coordinate.

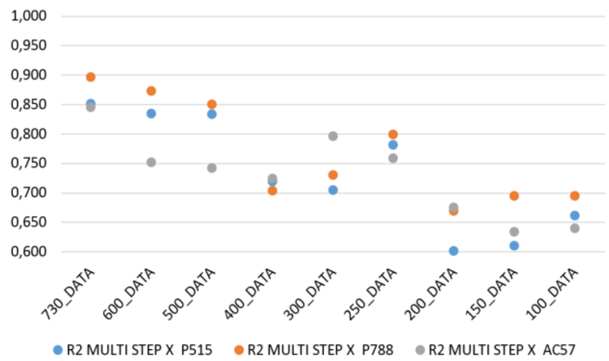

Figure 4. Results of R^2 for the multi-term forecast of the X (m) coordinate.

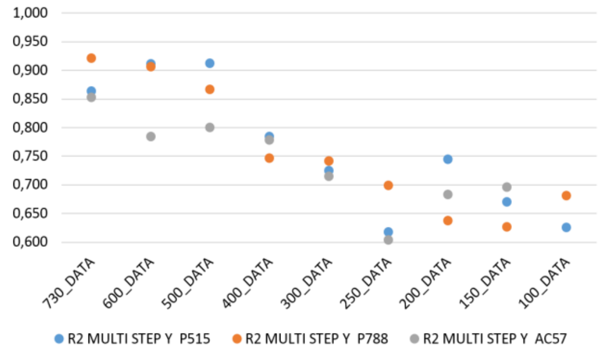

Figure 5. Results of R^2 for the multi-term forecast of the Y (m) coordinate.

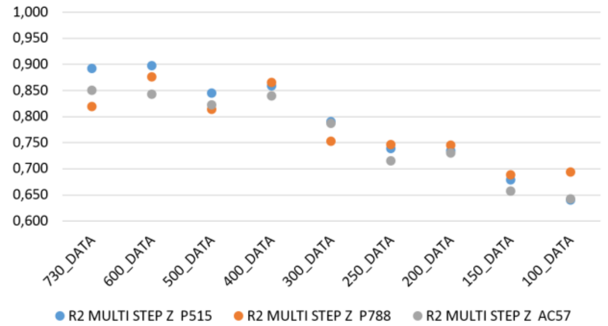

Figure 6. Results of R^2 for the multi-term forecast of the z (m) coordinate.

By the examination of the results obtained from the experiments, as far as the short-term forecast is concerned, the number of data from which the R^2 seems to be considerably reduced is the 250 daily data. Respectively, for the long-term forecast the same number is 300 daily records. The difference between the short-term and long-term is logical as long-term forecasting is by nature a more difficult process. The MAE of the three stations of the sample, which was calculated, is presented in the following table.

Table 2. MAE (mm) using optimum minimum inputs, for short-term forecasting and the position change dX, dY and dZ for the selected stratified random sample

Position	*MAE (mm) One-Step*			*Position Change (mm)*		
Changing	**P788**	**AC57**	**P515**	**P788**	**AC57**	**P515**
dX	0.9	1.0	1.2	1.9	1.6	1.9
dY	1.2	0.8	1.3	2.6	1.2	2.4
dZ	1.0	1.6	1.1	2.2	3.1	1.9

Table 3. MAE (mm) using optimum minimum inputs, for long-term forecasting and the position change dX, dY and dZ for the selected stratified random sample

Position	*MAE (mm) Multi-Step*			*Position Change (mm)*		
Changing	**P788**	**AC57**	**P515**	**P788**	**AC57**	**P515**
dX	1.5	1.1	1.2	2.0	1.8	1.9
dY	1.9	0.7	1.3	2.7	1.3	2.6
dZ	1.0	2.2	0.8	2.4	3.3	2.0

The research findings reveal that the number of inputs plays an important role during implementation of the methodology but at the same time, it is not infeasible to produce forecast with the desire accuracy with smaller amount of data.

IV. Conclusions & Future Research

This paper has described the ANNs-based methodology that is designed in order to produce accurate forecasts of 3D point position changing. The key word to every forecast research is the “accuracy”. The accuracy of the forecast is fundamental to many decision processes and hence the research for improving the proposed methodologies. However, the main idea was not only the application of this ANNs-based methodology with accurate results but to find the minimum necessary data needed in order to produce these kind of forecasts.

Therefore, in this paper, these two parameters were combined by examining the capability of artificial neural networks in forecasting 3D point position changing, when limited data are available.

This research aims to investigate the optimum minimum number, so that the produced short-term and long-term forecasts would have acceptable Mean Absolute Error (MAE), depending on the order of the individual position changing and an R^2 close to one.

The results show that we can produce reliable and accurate forecasts for 3D point position changing, even if big data are not available. Specifically, it is proven that there is indeed a minimum optimum number from which the forecasts cannot be considered as accurate. This number is 250 daily GNNS records for short-term forecasts and 300 for long-term.

After this research, some subjects that may be investigated in the future were emerged. In many geodetic applications related to detection of displacements, the used data are not continuous but come from phase measurements. Therefore, it is of particular interest to study the implementation of the methodology in such applications when data from phases are available (e.g. every month data, every year, etc.).

Another subject of future study is the use of a combined forecast method. This combination may include conventional and intelligent methods. For example, the use of fuzzy logic, support vector machines or neural networks with the ARIMA method can be investigated.

Finally, methods of transfer learning could be explored, i.e. methods that take advantage of a system's knowledge of a problem by transferring them to another related problem.

References

Abhishek, K., M.P. Singh, S. Ghosh, A. Anand (2012). Weather Forecasting Model using Artificial Neural Network. In: Procedia Technology, Vol. 4, pp. 311-318. DOI: 10.1016/j.protcy.2012.05.047

Alevizakou E.G. and G. Pantazis (2017). A comparative evaluation of various models for prediction of displacements. In: Applied Geomatics, Vol. 9(2). DOI: 10.1007/s12518-017-0189-8

Alevizakou, E.G, G. Siolas, G. Pantazis (2018). Short-Term and Long-Term forecasting for the 3D point position changing by using artificial neural networks. In: International Journal of Geo-Information, Vol. 7, 86

Allende, H., C. Moraga and R. Salas (2002). Artificial neural networks in time series forecasting: a comparative analysis. In: Kybernetika, Vol. 38, No.6, pp. 685-707

Babu, B.V., V. Ramakrishna, K. Chakravarthy (2003). Artificial Neural Networks for Modeling of Adsorption. In: Proc of the Second International Conference on Computational Intelligence, Robotics, and Autonomous Systems (CIRAS-2003), Singapore, 15–18 December

Bell, J.W., F. Amelung, A.R. Ramelli, G. Blewitt (2002). Land subsidence in Las Vegas, Nevada, 1935–2000: New geodetic data show evolution, revised spatial patterns and reduced rates. In: Environ. Eng. Geosci., 8, pp. 155–174. DOI: 10.2113/8.3.155

Cai, S., X. Kang and L. Zhang (2012). Time Series Prediction Based on Artificial Neutral Network for Estimation of Forest Biomass. In: International Journal of Advancements in Computing Technology (IJACT), Vol. 4, No.12, July 2012.

Chan, W.S., Y.L. Xu, X.L. Ding, Y.L Xiong, W.J. Dai (2005) Dynamic displacement measurement accuracy of GPS for monitoring large civil engineering structures. In: Proceedings of the SPIE, International Symposium on Smart Structures and Nondestructive Evaluation (CDROM), San Diego, CA, USA, 17 May 2005

Demir, A., A. Shadmanov, C. Aydinli, O. Eray (2015). Designing a forecast model for economic growth of Japan using competitive (hybrid ANN vs multiple regression) models. In: *Ecoforum*, Vol. 4, Issue 2(7)

Elsayir A.H. (2014). Comparison of precision of systematic sampling with some other probability sampling, In: American Journal of Theoretical and Applied Statistics, Vol. 3, Issue 4, pp. 111-116. DOI: http://dx.doi.org/10.11648/j.ajtas.20140304.16

Erdogan S. (2010). Modelling the spatial distribution of DEM error with geographically weighted regression: An experimental study. In: Computers and Geosciences, Vol. 36, pp. 34–43

Farmer, G.T., A.V. Newman, P. Psimoulis, S. Stiros (2007). Geodetic characterization of Santorini Caldera from GPS measurements. In: Proceedings of the AGU-2007 Fall Meeting, San Francisco, CA, USA, 10–14 December 2007.

Kareem, A., I. Oshungade and G. Oyeyemi (2005). A note on the Precision of Stratified Systematic Sampling In: Open Journal of Statistics. DOI: http://dx.doi.org/10.4236/ojs.2015.52013

Khalil, B.M., A.G. Awadallah, H. Karaman, A. El-Sayed (2012). Application of Artificial Neural Networks for the Prediction of Water Quality Variables in the Nile Delta. In: Journal of water resource and protection, Vol. 4, No. 6, pp. 388–394. DOI: 10.4236/jwarp.2012.46044

Mounce S.R. (2013). A comparative study of artificial neural network architectures for time series prediction of water distribution system flow data. University of Exeter, 3rd-5th April, 2013.

Moshiri, S. and N. Cameron (2000). Neural Network versus econometric models in forecasting inflation. In: *Journal of Forecasting,* Vol. 19, pp. 201-2017

Psimoulis, P., M. Ghilardi, E. Fouache, S. Stiros (2007). Subsidence and evolution of the Thessaloniki plain, Greece, based on historical leveling and GPS data. In: Eng. Geol. 2007, Vol. 90, Issues 1-2, pp. 55–70. https://doi.org/10.1016/j.enggeo.2006.12.001

Safavieh, E., S. Andalib and A. Andalib (2007). Forecasting the Unknown Dynamics in NN3 Database Using a Nonlinear Autoregressive Recurrent Neural Network. In: Sepanta Robotics & AI Research Foundation (SRRF)

Schroeder M., D. Cornford, I. Nabney (2009). Data visualization and exploration with prior knowledge. In: Engineering applications of neural networks, Vol. 43. Springer, pp. 131–142

Yilmaz M. and M. Gullu (2014). A comparative study for the estimation of geodetic point velocity by artificial neural networks. In: J. Earth Syst. Sci., Vol. 123, No. 4, pp. 1–18

EarthScope - Data. Available online: http://www.earthscope.org/node/395 (accessed on November 2018).

EarthScope - Plate Boundary Observatory (PBO). Available online: http://www.unavco.org/projects/majorprojects/pbo/pbo.html (accessed on November 2018).