GEOCHRONOMETRIA Vol. 20, pp 45-50, 2001 – *Journal on Methods and Applications of Absolute Chronology*

NONLINEAR METHODS OF ANALYSIS OF DATA WITH GAPS

V. A. DERGACHEV[1], N. G.MAKARENKO[2], L. N. KARIMOVA[2] and E. B. DANILKINA[2]

[1]Ioffe Physico-Technical Institute, Politekhnicheskaya 26, 194021 St.Petersburg, Russia
(e-mail: v.dergachev@pop.ioffe.rssi.ru)
[2]Institute of Mathematics, Almaty, 4801100 Kazakstan

Key words:
A NEURAL NETWORK,
ATTRACTOR,
NATURAL TIME SERIES,
MODELING,
RADIOCARBON CONCENTRATION,
TREE-RING WIDTH,
WOLF NUMBER,
SEA LEVEL,
BERILLIUM-10 CONTENT

Abstract: **Information on most of natural phenomena can be obtained from time series of direct and proxy data. The analysis of time series generated by natural dynamic systems is a key element in interpreting geophysical and climatic information. Unfortunately, most of available time series have gaps. When there are many gaps with irregular distribution, we do not have any statistical tools for repairing the data. We suggest some approach to solve this problem. It is based on modeling the missing data by small-dimensional manifolds and neural network technologies. In this approach we assume that data under consideration are a set of *n*-dimensional vectors, which are produced by dynamical system. These vectors model *n*-dimensional attractor in embedding space. Gaps in the vectors are represented as a linear manifold *L* of some dimension. The method idea is to model *L* by another small-dimensional manifold, e.g. a curve. Neural networks are used to find this manifold. We verify the method on real time series data: sunspot numbers, the radiocarbon content in tree rings, the ^{10}Be in ice cores, the width of tree rings and so on.**

1. INTRODUCTION

By the present time large experimental data about geophysical, astrophysical, biological phenomena are accumulated. These data are represented in the form of time series. The objective of this data analysis is to describe and to elucidate the nature of underlying physical processes producing these time series. Information carried in a time series is frequently a superposition of different natural processes with different scales of coherence or memory. Besides, records are often contaminated with noise. All these factors lead to complex nonlinear and non-stationary structure of the data. For time series produced by chaotic dynamic systems of small dimension, there are certain quantities, e.g. attractor dimension or Lyapunov exponents, which can be obtained with the help of modern topology tools (Sauer, Yorke and Casdagli, 1991). These quantities are of special interest, because they characterize intuitively useful concepts of underlying physical systems, e.g. number of active degrees or its predictability. However, in order to apply this technique one should have quite long and equidistant data. Practically, available time series have short length, part of them can be lost, and these series are non-equidistant. This fact is a substantial obstacle for obtaining reliable results. Traditional methods of gaps recovering are not effective for non-stationary and nonlinear time series (Little and Rubin, 1987). We suggest an approach to solve the problem of recovering the missing data based on chaotic dynamics methods and neuromathematics (Danilkina, Kuandykov and Makarenko, 2001; Dergachev *et al.*, 2001). The paper structure is as follows. In section 1 we recall some ideas of algorithmic data modeling. In section 2 the method of data recovering is described. In section 3 we give the results of the method application to real data.

2. ALGORITHMIC MODELING OF DATA TABLE

To explain the main idea, let us begin with the table structure. Let us have original data as discrete finite time series $\{x_i\}_{i=1}^{N}$. Assume that this series is nonlinear projection of an orbit of some unknown dissipative dynamic system onto an arbitrary direction. Also this dynamic system is supposed to have an attractor of finite dimension *d* in its phase space. In other words, the orbits are attracted to some compact small-dimensional region as $t \to \infty$. Then under some considerations one can reconstruct this system attractor copy as topological embedding the time series to Euclid space of an appropriate dimension $m \geq 2d + 1$ (Sauer, Yorke and Casdagli, 1991).

Dynamics of the unknown system in this space is a map of shift of *m*-dimensional vectors:

$$(x_i, x_{i+1}, ..., x_{i+m-1}) \to (x_{i+1}, x_{i+2}, ..., x_{i+m}) .$$

Such a shifts sequence forms an orbit of our model which is the true attractor copy. It is known that this copy typically preserves all dynamical invariant (attractor dimen-

sion, Lyapunov exponents etc.) of the real system up to continuos nonlinear maps (diffeomorphisms) (Sauer, Yorke and Casdagli, 1991; Eckmann and Ruelle, 1985).

As long as we do not know the embedding dimension m, one should try to test dimensions $m = 2,3,...$ For each of test dimensions one should construct a copy of m-dimensional attractor and to obtain some measure estimation of it. This measure can be as follows: number of nonempty cubes, necessary for covering all the points of the attractor, or number of ε-close pairs of vectors. These measures change with m. But as we reach the correct value of m, the measure is stabilized. The simplest way to determine m is the false neighbor method (Kennel, Brown, and Abarbanel, 1992). For each point of the attractor copy in R^m let us determine ε-neighborhood, which contains ε-close points. When m increases, some part of these neighbors are found to be false – they leave the ε-neighborhood when we go to $m + 1$. Then the true dimension m is that, for which the number of false neighbors does not change for $m + 1$.

If the time series has missed values, then one can use long fragments of the series without gaps for these procedures. If we do not have such fragments, one can *temporarily* substitute the missed values for mean values because the attractor copy is reconstructed only up to diffeomorphism. So, suppose that we have determined the dimension m.

Let us construct a table of data where each row is m-dimensional vector of the attractor copy. A sequence of the rows is a sequence of points of our model orbit. Then in the table an absent diagonal corresponds to each missed value. For example, for $m = 5$ and x_4 missed, we have the table

$$x_1, x_2, x_3, \varnothing, x_5$$
$$x_2, x_3, \varnothing, x_5, x_6$$
$$x_3, \varnothing, x_5, x_6, x_7$$
$$\varnothing, x_5, x_6, x_7, x_8$$
$$\ldots\ldots\ldots\ldots\ldots$$

where ϕ indicates the missed data. Let us further denote such a table as a matrix $\mathbf{A} = \left(a_{ij}\right)$.

3. THE METHOD OF DATA RECOVERING

Let some of the rows have $k<m$ gaps. The gaps of the vector are considered as k-dimensional linear manifold L_x parallel to k coordinate axes. Thus, initial data are figured as a set of the points and the gaps – linear manifolds of space R^m. The problem of gaps recovering comes to a search of some manifold M of a given small dimension, which approximates the data in the best way (Gorban, Rossiev and Wunsch, 2000). For complete vectors of data the accuracy of approximation is defined as an ordinary distance from a point to M (lower bound of the distances to the set points). For incomplete rows the minimum of all distances between the points of the sets M and L_x is used.

The method has three versions: linear, quasi-linear and nonlinear (Eckmann and Ruelle, 1985; Gorban, Rossiev and Wunsch, 2000). They differ in the way of construction of the manifold M. In the linear method the data are modeling by a linear manifolds sequence of small dimension. That is the initial matrix **A** is approximated by the sequence of the matrices $P_1, P_2, \ldots, P_q$ of a form $\{z_i y_j + b_i\}$, where y_i and b_i – components of some vectors, q is a number of factors. For the matrix **A** one looks for the best approximation by the matrix $\mathbf{P_1}$. Analogously, for the matrix $\mathbf{A\text{-}P_1}$ the approximation $\mathbf{P_2}$ is looked for and so on. The stopping criterion is sufficient approximation to zero of the matrix elements calculated at every iteration. As a result we obtain a singular expansion of the matrix **A** presented by the sum of matrices $\mathbf{A = P_1 + P_2 + ... + P_q}$. Q-factor gaps recovering lies in their finding from the sum of the obtained matrices P_i, $i = 1, 2, \ldots, q$. Q-factor „repair" of the matrix **A** is its replacement by the sum of the matrices P_i, $i = 1, 2, \ldots, q$.

Geometrical interpretation of the linear (a) and quasi-linear (b) models in R^3 are presented in **Fig. 1**. Let one of the rows of the matrix **A** look like a $\{\phi, \phi, x_3\}$ then the plane L_k can be formally represented as a scheme $\{\bullet, \bullet, x_3\}$, where bold dots indicate any values from R. So L_k passes through the point $\{0, 0, x_3\}$ and it is parallel to the plane formed by the axes x_1 and x_2. In the case of the linear model initial set of points is modeled by manifold y, which is a sloping plane, approximating the known data in the best way (vectors **z** are a collection of projections of the initial data onto $y \equiv M$). The intersection $L_k \cap M$ in the course of successive approximation gives the recovered value. In the case of the quasi-linear model (b) a curve $f(t)$ is constructed instead of the sloping plane. The $f(t)$ must be a nonlinear piecewise smooth function.

The recovered value is a point of intersection $f(t) \cap L_k$.

As indicated in the picture, the quasi-linear model is similar to the linear model, except inputting the certain vector-function $f(t)$ only. So, at first, the linear model is constructed and thereafter the vector-function $f(t)$ minimizing the functional Φ is found.

$$\Phi = \sum_{\substack{i,j \\ a_{ij}\neq\varnothing}} \left(a_{ij} - f_j\left(\sum_k a_{ik} y_k \right) \right)^2 + \alpha \int_{-\infty}^{+\infty} \left(f''(t) \right)^2 dt,$$

where $\alpha > 0$ is a smoothing parameter.

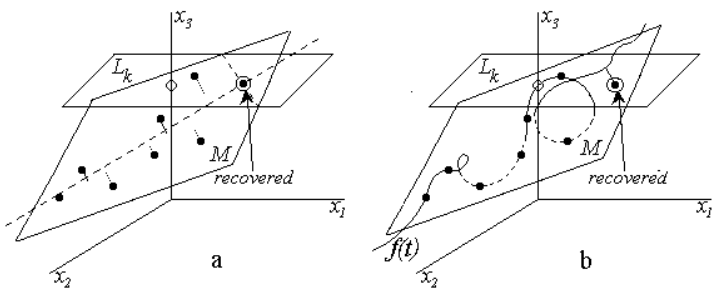

Fig. 1. *Geometrical interpretation of the linear model (a) and the quasi-linear (b) model.*

In the nonlinear variant of the method the data are modeled by self-organizing curves (Gorban, Rossiev and Wunsch, 2000). Practical realization of the method is a neural package Famaster-2 made in the Computing Modeling Institute (Krasnoyarsk, Russian) by (Gorban and Rossiev, 2000).

4. EXPERIMENTAL RESULTS

Below are given the results of recovering gaps in different time series. For all cases the optimal value of embedding dimension was defined by false neighbor method (Kennel, Brown and Abarbanel, 1992).

Diagnostics of the method was performed in the following way: at the beginning for the initial time series the embedding dimension was determined and then the table was constructed. Then a part of the data (5-50%) was randomly deleted and subsequently the recovering of the data was performed. As long as the diagonal in the table corresponds to one missed value, the mean value of the diagonal as the most plausible estimation was used. Different time series were used for the experiments.

Example 1: the ^{10}Be concentration time series. Fig. 2 and Table 1 show the results obtained for the cosmogenic isotope ^{10}Be time series (Beer *et al*., 1994). It should be emphasized that the initial time series has 5% of *real* gaps.

Table 1. *The recovering of missed values in the time series of the ^{10}Be concentration. In the left columns there are actual values, some of them are absent, in the middle ones – recovered values of the absent data, the right columns contain recovered values which were eliminated by hand.*

Actual values	Recovered real	Test	Actual values	Recovered real	Test
0.6600	-	0.6706	-	0.8986	-
-	0.7598	-	-	1.1757	-
0.4100	-	0.4499		0.8361	-
1.0300	-	1.0174	-	1.1031	-
0.5700	-	0.8447	-	1.1549	-
-	0.7395	-	-	1.2067	-
1.1400	-	0.9658	-	0.8561	-
-	1.0032	-	-	0.9921	-
-	1.0306	-	-	0.8970	-
-	1.1845	-	2.0795	-	0.8894
1.1900	-	1.1234	0.8233	-	0.7381
0.7800	-	0.7700	-	1.3028	-
-	0.8632	-	-	1.1884	-
0.6059	-	0.8338	1.1981	-	0.7462
1.5635	-	1.2474	0.6988	-	1.1613
1.0898	-	0.9918	1.5633	-	1.4373
-	0.9672	-	1.3255	-	1.2029
0.6863	-	0.9945	0.7189	-	0.7108
-	0.9167	-	0.8711	-	1.1416
0.4602	-	0.7127	-	1.4030	-
-	0.9368	-	-	1.3692	-
-	1.5939	-	-	1.2878	-
-	1.6752	-	-	1.4084	-
-	1.2345	-	-	1.3010	-
0.9017	-	1.0120	-	1.1316	-
-	1.0097	-	-	1.2200	-

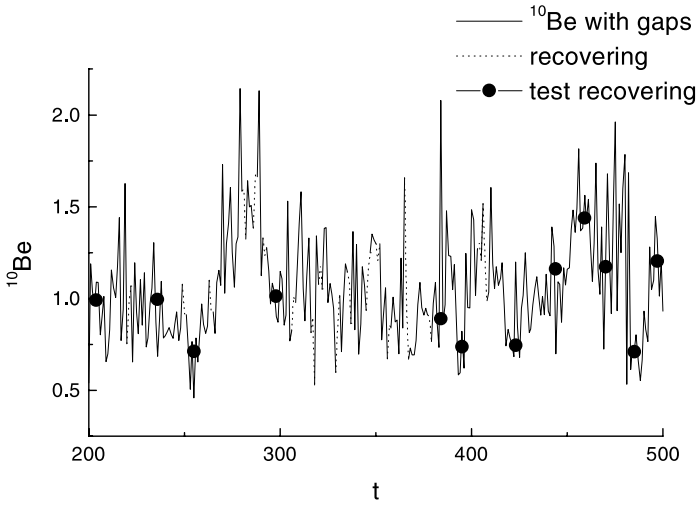

Fig. 2. *The results of recovering missed points (10%) in a part of initial time series of the ^{10}Be concentration in time interval 1428-1985 years by the quasi-linear model and $m = 6$.*

Before the procedure of recovering data, we randomly deleted 5% of the points (in addition to real ones) to test the degree of certainty of quasi-linear model operation. The averaged diagonal values in the table were used to close the gaps. The table was constructed in accordance with the dimension of embedding $m = 6$, i.e. the number of columns was also 6.

Example 2: **tree rings time series**. **Fig. 3** shows the results of recovering values of the Juniperus turkestanica tree rings on the upper boundary of forest (Lovelius, 1997) during a period of 1163-1970 years. From the initial time series including 808 point about 10% of points were randomly deleted. The recovering was carried out by the SOC-model (self-organizing curves) with $m = 7$.

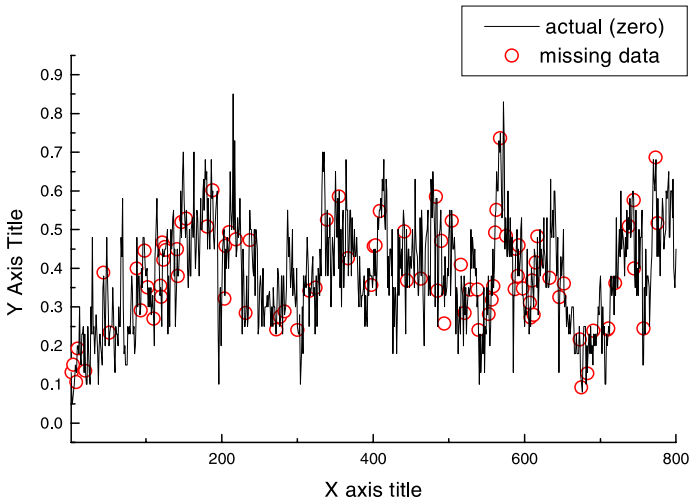

Fig. 3. *The results of recovering data in the time series of tree-ring width after deleting ~10% points using the SOC-model with $m = 7$.*

Example 3: The Caspian Sea level data. The annual data of the Caspian Sea level (Sadykov *et al*. 1995) were used with $m = 6$ and SOC model. Approximately 10% of points were randomly eliminated. The results of the recovering data on the time interval 1830-1999 years are shown in **Table 2** and **Fig. 4**.

Table 2. *The results of estimation of the recovering values of data indicated by circles in Fig. 4.*

Gap's number	Real value	Recovered value	Gap's number	Real value	Recovered value
14	-26.30	-26.2	99	-26.11	-26.11
15	-26.22	-26.19	107	-26.78	-26.81
23	-26.27	-26.29	108	-26.99	-27.06
29	-26.4	-26.37	126	-28.36	-28.36
52	-25.51	-25.64	128	-28.38	-28.34
53	-25.4	-25.46	135	-28.37	-28.39
55	-25.71	-25.71	144	-28.6	-28.64
73	-25.87	-25.79	149	-28.95	-28.85
84	-26.34	-26.27	152	-28.25	-28.25
85	-25.23	-25.87	156	-27.96	-27.97
86	-25.99	-25.95	161	-27.53	-27.46
95	-26.45	-26.45	166	-26.2	-26.66

Let us denote, that using the method described above the time series of annual data could be extended to the past, before 1830 year, on the base of available non-equidistant data, which are used only after aggregation in 3 or 5 years data.

Example 4: radiocarbon concentration time series. Here we increased the number of deleted points. **Fig. 5** shows the results of recovering 30% of the points in the time series of the radiocarbon concentration in the Earth's atmosphere over time interval from 5995 BC to AD 10 (Stuiver and Becker, 1993).

Example 5: annual Wolf number index. About 50% of points were randomly deleted from the time series of the number of sunspots. **Fig. 6** shows the results of gaps recovering with SOC model and $m = 6$.

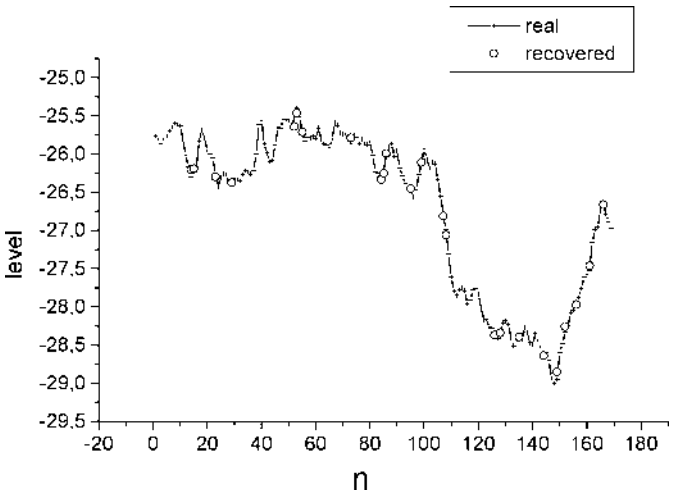

Fig.4. The results of modeling the Caspian Sea level changes during time interval 1830-1999 years.

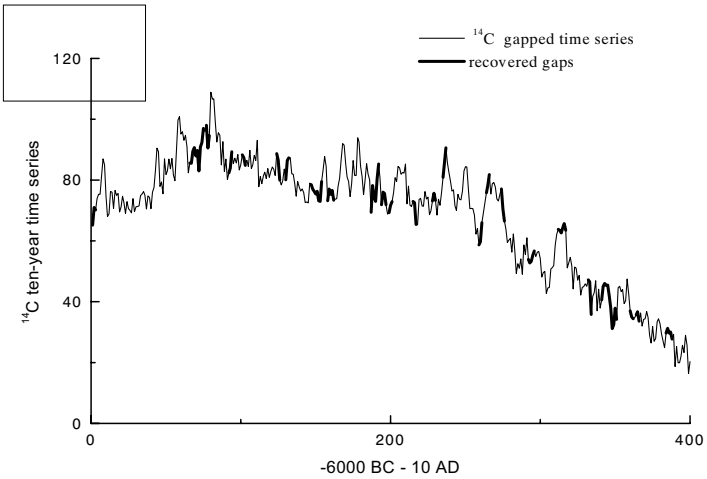

Fig. 5*. The results of recovering gaps in the ^{14}C concentration after deleting 30% points in the initial time series with the quasi-linear model and m = 7.*

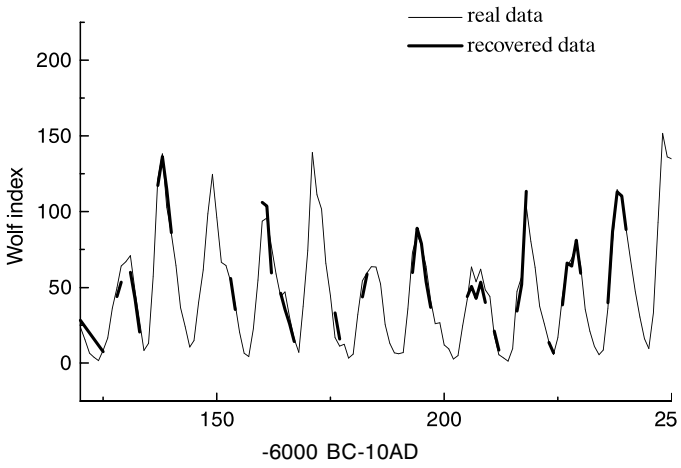

Fig.6. *Modeling gaps in annual Wolf numbers after deleting 50% of points. A fragment for the period 1820-1950 is given. SOC model, m = 6.*

5. CONCLUSIONS

Numerical experiments on different time series have shown an effectiveness of the proposed method. It should be emphasized that a) this method gives only the most plausible values, not the most probable, because there was not made any assumption about the underlying statistics and this method is found on the Anzatz-hypothesis; b) it is successful only when the table is constructed according to Takens algorithm (Sauer et al., 1991). However, if number of gaps is significant, and their distribution is irregular, other methods are not applicable at all. Our tests have demonstrated that the approach described above can be successfully applied to historical data preprocessing. Besides, it is irreplaceable for data repair, i.e. correcting values, which raise doubts for some reasons.

ACKNOWLEDGEMENTS

This work was supported by the INTAS grant 97-31008.

REFERENCES

Beer J., Baumgartner St., Dittrich-Hannen B., Hauenstein J., Kubik P., Lukasczyk Ch., Mende W., Stellmacher R. and Suter M., **1994:** Solar variability traced by cosmogenic isotopes. In: Pap J.M., Frohlich C., Hudson H.S. and Solanki S.K., eds., *The Sun as a Variable Star: Solar and Stellar Irradiance Variations,* Cambridge University Press: 291-300.

Danilkina E.B., Kuandykov Y.B. and Makarenko N.G., 2001: The neural networks and chaos in the problem of non-complete data reconstruction (in Russian). *Neuroinformatics-2001*, Moscow, MIFI: 166-173.

Dergachev V.A., Makarenko N.G., Kuandykov E.B. and Rossiev A.A., **2001:** How to discover interrelation between two dynamical systems using observed time series with gaps (in Russian). Izvestiya Academii Nauk, Seriya Fizicheskaya 65(3): 391-393.

Eckmann J.P. and Ruelle D., **1985**: Ergodic theory of chaos and strange attractors. *Review of Modern Physics* 57(3): 617-656.

Gorban A.A., Rossiev A.A. and Wunsch II D.C., 2000: Self-organizing curves and modeling by neural networks of data with gaps (in Russian). *Neuroinformatics-2000*, Moscow, MIFI: 40-46.

Gorban A.A. and Rossiev A.A., 2000: Iterative Modeling Data with Gaps by Self-organizing Small-dimensional Manifolds with the help of Neural Networks (in Russian). *Neuroinformatics and its Applications*, Proceedings of VIII Seminar, October 6-8, 2000, Krasnoyarsk: 45-48.

Kennel M.B., Brown R. and Abarbanel H.D.I., **1992**: Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review* A 45: 3403 -3411.

Little R.J.A. and D.B. Rubin., 1987: *Statistical Analysis with Missing Data*. John Willey and Sons.

Lovelius N.V., 1997: *Dendroindication of natural processes and anthropogenic influences*. World and Family-95, St.Petersburg: 320 p.

Sadykov Zh.S., Golubtsov V.V. and Duisenbaev Zh.D., 1995: Changes of the Caspian Sea level and its prediction (in Russian). *Doklady Academy Nauk Kazakhstan Respublika* 6: 9-19.

Sauer T., Yorke J.A. and Casdagli M., 1991: Embedology. *Journal of Statistical Physics* 65(3/4): 579-616.

Stuiver M. and Becker B., 1993: High precision decadal calibration of the radiocarbon time scale AD 1950-6000 BC. *Radiocarbon* 35 (1): 35-65.