THE FILLING OF GAPS IN GEOPHYSICAL TIME SERIES BY ARTIFICIAL NEURAL NETWORKS

V A Dergachev[1] • A N Gorban[2] • A A Rossiev[2] • L M Karimova[3] • E B Kuandykov[3]
N G Makarenko[3] • P Steier[4]

ABSTRACT. Nowadays, there is a large number of time series of natural data to study geophysical and astrophysical phenomena and their characteristics. However, short length and data gaps pose a substantial problem for obtaining results on properties of the underlying physical phenomena with existing algorithms. Using only an equidistant subset of the data with coarse steps leads to loss of information. We present a method to recover missing data in time series. The approach is based on modeling the time series with manifolds of small dimension, and it is implemented with the help of neural networks. We applied this approach to real data on cosmogenic isotopes, demonstrating that it could successfully repair gaps where data was purposely left out. Multi-fractal analysis was applied to a true radiocarbon time series after recovering missing data.

INTRODUCTION

The analysis of time series generated by natural dynamic systems has been a key element in interpreting geophysical and climatic information. The goal of this analysis is to describe and elucidate the nature of the underlying physical processes producing a signal in the data under consideration. The problem of extracting information from dynamic systems is very important in many scientific and practical areas. In this paper, we focus on time series of cosmogenic isotopes.

Information carried in a time series is frequently a superposition of different processes with different scales of coherence or memory. In many cases the correlation structure of the time series X(t) shows the property of stochastic self-similarity: A magnified small section looks similar to a large one over a wide range of scales. For stochastic objects, self-similarity is used in the distributional sense: when viewed at varying scales, the object's stochastic distributions remain unchanged. Time series showing stochastic self-similarity are often considered to be multi-fractal processes that can be described with the help of scaling exponents (Falconer 1994), by characterizing their singularities (MacDonald 1989; Davis et al. 1994a), and by identifying irregular structures (Davis et al. 1994b). Real natural records are contaminated with noise, which is rarely additive, Gaussian, or white. Often, noisy structure of data is generated by underlying non-linear chaotic processes (Abarbanel et al. 1993b), which exhibits alternating periodic, quasi periodic, and chaotic patterns.

These factors lead to complex-structured non-linear and non-stationary properties. For time series arising from low-dimension chaotic systems, there are certain quantities, e.g. the dimension of the attractor or the Lyapunov exponents, that can be obtained using up-to-date topology tools (Abarbanel et al. 1993a; Sauer et al. 1991). These quantities are especially interesting, as they characterize intuitively useful concepts of the underlying physical systems, e.g. its number of active degrees of freedom, or its predictability. If we cannot assume the existence of underlying low dimension dynamics, we can use the Wold decomposition (Anderson 1971), which is satisfied for any stationary process.

Thus, nowadays investigators have a few tools to analyze complex data. Nevertheless, to apply these techniques one should have long enough and equidistant time series. Considering the short length of

[1] Cosmic Ray Laboratory, Ioffe Physico-Technical Institute, St. Petersburg 194021, Russia.
Email: v.dergachev@pop.ioffe.rssi.ru.
[2] Institute of Computational Modeling SD RAS, Akademgorodok, Krasnoyarsk-36 660036, Russia
[3] Institute of Mathematics, Almaty 480100, Kazakhstan
[4] VERA Laboratorium, Institut für Isotopenforschung und Kernphysik Universität Wien, A-1090 Wien, Austria

RADIOCARBON, Vol 43, Nr 2A, 2001, p 365–371
Proceedings of the 17th International ^{14}C Conference, edited by I Carmi and E Boaretto

https://doi.org/10.1017/S0033822200038224 Published online by Cambridge University Press

available cosmogenic time series, in particular if some part of the data is lost, this is a substantial problem for obtaining reliable results. Non-equidistant data distort even ordinary statistic characteristics, and using only an equidistant subset with coarse steps leads to a loss of information. Traditional methods of filling gaps are not effective for non-stationary and non-linear time series (Little and Ruben 1987). In the case of a lot of missing data and when its location is random there is no known solution. We suggest an approach to solve the problem of the recovery of missing data in time series using artificial neural networks.

Recovering Gaps in the Data by Neuromathematical Methods

Available time series have different indices, which are often heterogeneous: they may have a different length, different argument steps, and often some fragments are lost. We approach the problem of recovering gaps in time series using a new neural non-linear method (Rossiev 1998; Gorban et al. 1998) of modeling data with gaps by a sequence of curves. The method is a generalization of the iterative construction of the singular expansion of matrices with gaps (Hastie and Stuetzle 1988; Kramer 1991). The method itself is founded on Ansatz-reasoning and only allows one to obtain the plausible values of the missing data. However, the testing by means of artificially gapped time series has shown remarkable results.

The Basic Model

The idea of modeling data with the help of manifolds of small dimension was conceived long ago. Its most widespread, old, and feasible implementation for data without gaps is the classical method of principal components. The method calls for modeling the data by their orthogonal projections over "principal components". Generally, to present data with sufficient accuracy requires relatively few principal components.

We assume that our data are a set of n-dimensional vectors, which form the rows of a table $\mathbf{A}=\{a_{ij}\}$. Let a part of information in the table be missing - there are some gaps a_{ij}=@ for some i, j (the symbol @ is used to denote gaps in the data). Let us take a look at the row **x** which forms a vector (x_1, x_2,...,x_n). There may be k gaps in the vector **x**, i.e. some of the components of **x** are unknown. Therefore this vector represents a k-dimensional linear manifold L_x, parallel to k coordinate axes corresponding to the missing data. If there are additional a-priori restrictions on the missing values, instead of L_x we obtain a rectangular parallelepiped $Q_x \subset L_x$.

We search for a manifold M of a given small dimension (in most cases a curve, dimension = 1) approximating the data set in the best way and satisfying certain additional regularity conditions. The quality of the approximation is determined from the lower bound of the distances between the points of M and L_x, (or, accordingly, Q_x). We obtain a residual by subtracting from each data vector the closest point of M. The process is repeated until the residuals are close enough to zero. Approximations can be constructed recursively (Rossiev 1998; Gorban et. al. 2000) for three models: linear, quasi-linear or essentially non-linear (self-organizing curves—SOC). We illustrate the idea starting with a linear model as an example.

In the first step we approximate the data **A** with a straight line M. The points on **M** which are closest to the data points form a matrix $\mathbf{P_1} = \{c_i y_j + b_j\}$. The vectors **y** and **b** are found by minimizing the quality of the approximation, which is measured by means of the least-squares method:

(1)

Further, we are looking for a matrix P_2 that is the best approximation of the residual $\mathbf{A}$-$\mathbf{P}_l$ and so on, while the norm of the residual isn't sufficiently close to zero. Thus, the initial matrix **A** is presented in a form of a sum of **q** matrices, i.e. $\mathbf{A}\approx\mathbf{P}_1+\mathbf{P}_2+...+\mathbf{P}_q$. The q-factorial recovering of the gaps consists in their definition through the sum matrix. For incomplete data, with a number of iterations we get a system of factors which we will use for recovering.

Figure 1a shows the geometrical interpretation of a linear model, for x(**i**) $\subset \mathbf{R}^2$, i = 1,2,…. We assume that for one point **x** one coordinate is lost, so it corresponds to a line L_x. The data is approximated by the inclined line **M** (with direction parallel **y**), which approximates the known data in a best manner. The matrix $\mathbf{P}_1$ consists of the points on **M** closest to the initial data set. The lost coordinate of **x** is substituted by the intersection $\mathbf{L}_x \cap \mathbf{M}$.

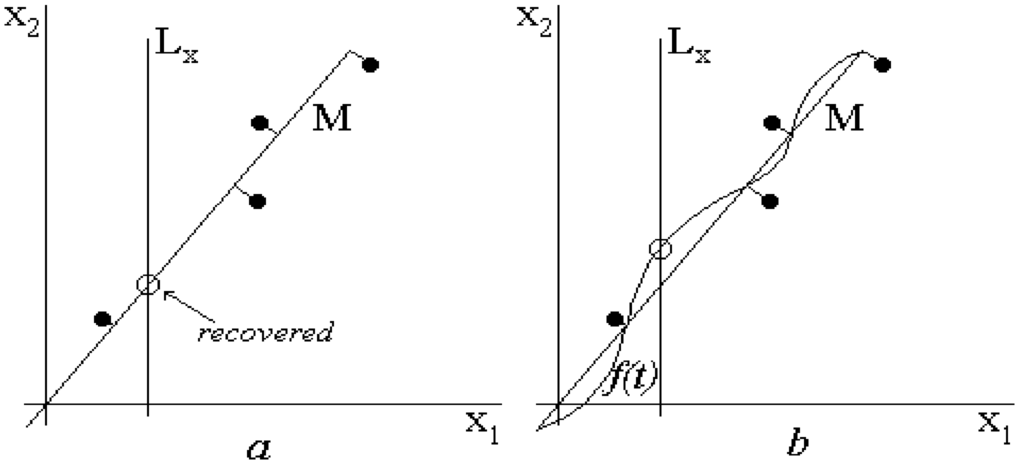

Figure 1 Geometrical interpretation of the linear model (a) and the quasi-linear model (b).

Quasi-linear models (Rossiev 1998; Gorban et. al. 2000) are constructed in several stages: First, we construct a linear model with the algorithm explained above, yielding the vectors **y** and **b**. These vectors can always be chosen so that $\mathbf{y}\cdot\mathbf{b} = \mathbf{0}$ and $\mathbf{y}\cdot\ \mathbf{y}=1$.

Then, we construct a vector-function $\mathbf{f}(t) = [f_1(t),...f_n(t)]$ (a cubic spline or a polynomial) which minimizes the function

$$\Phi = \sum_i \left|\mathbf{a}_i - \mathbf{f}(\mathbf{a}_i \cdot \mathbf{y})\right|^2 + \alpha \int_{-\infty}^{+\infty} \left|\mathbf{f}''(\tau)\right|^2 d\tau \tag{2}$$

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

So, first we are looking for the projection of data vector **a** on the line **y:** Pr(**a**)=t**y**+b, t=(**a**,**y**), then we find a point on the curve *f*(t). For incomplete data it is taken the closest point t(**a**) on the line. And after that we take the corresponding point on the curve f(t) for t=t(**a**). After construction of f(t) the matrix **A** is substituted by the matrix of deviations from the model (see Figure 1b). The process is repeated several times and in the end the initial table **A** is represented in the form of the q-factorial model: $a_{ij} \cdot \approx \Sigma_{i,j} f_i(t_j)$.

https://doi.org/10.1017/S0033822200038224 Published online by Cambridge University Press

The third model is based on the theory of Kohonen self-organizing maps (or, more precisely, on the paradigm of self-organizing curves). These curves are defined by a set of model points (a kernel). In the first approximation polygons are used. Every data point is mapped onto the closest point of the kernel. The domain of points mapped to a certain kernel point is called its taxon. The kernel points are to be placed in a way that 1) the total length of the curve is minimal, 2) the summed distance of the data points from their respective taxons is minimal, and 3) the angles between adjacent segments of the polygon is minimal (Gorban et. al. 2000). This can be achieved iteratively: under fixed decomposition of the data set into taxons the kernel points are calculated. Under fixed location of kernels, the taxons are re-determined. Successive searching of kernels → taxons → kernels → … leads to a convergent algorithm. The final smoothening of the polygon is done analogous to the method used in the quasi-linear model.

The computational process is implemented on the neural conveyor Famaster-2 made by Gorban's team at the Institute of Computational Mathematics of SD RAS (Russia). Contrary to the original approach (Rossiev 1998; Gorban et. al. 2000), in this paper we form a data table according to Takens algorithm (Sauer et al. 1991). So, we assume that the data are produced by the underlying dynamic system, whose trajectories are continuous and belong to low dimensional attractor.

RESULTS

We carried out experiments with different time series. Figure 2 shows the results obtained for the annual Wolf index time series after deleting about 50% of the points. The gaps were recovered using the SOC model. Vectors of the initial data table were Takens' m-dimensional retarded vectors (Sauer et al. 1991) with delay 1 and embedding dimension 6, i.e. the k-th vector consists of 6 consecutive elements starting with the k-th element in the time series: $\mathbf{a}_k=(x_k, x_{k+1},..., x_{k+5})$. Therefore, a missing point in the time series implies missing elements in 6 data vectors. As can be seen in Figure 2, the neural conveyor remarkably recovered even the peaks of the cycles.

Figure 3 shows a part of the cosmogenic isotope ^{14}C time series (Stuiver and Becker 1993). We show the results for deleting and recovering 30% of the points in the time range from 5995 BC to 10 AD. The last time series we used in our experiment was a ^{10}Be time series (Beer et al. 1994), where we deleted about 10% of all points (Figure 4).

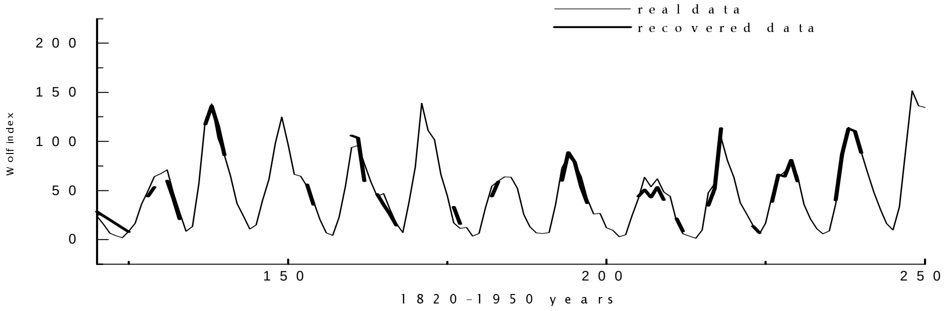

Figure 2 A fragment of the annual Wolf index time series, SOC, number of mode: 10

https://doi.org/10.1017/S0033822200038224 Published online by Cambridge University Press

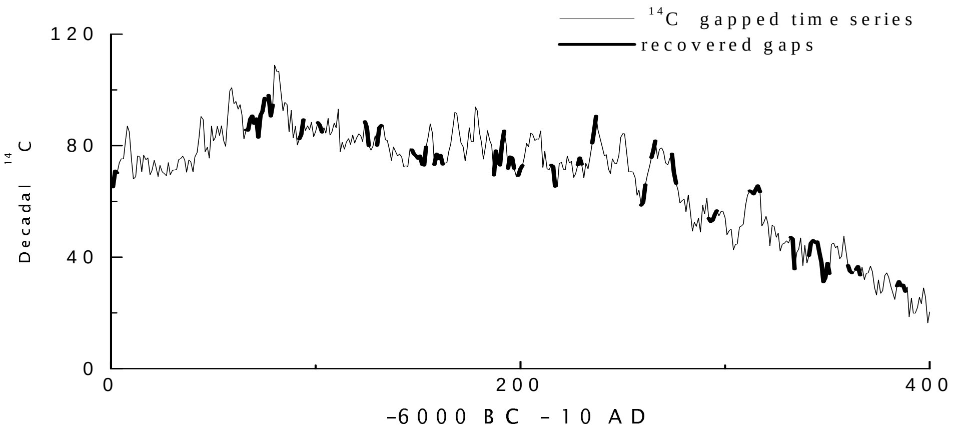

Figure 3 A fragment of the ^{14}C 10-year time series, quasi-linear model, number of nodes: 8

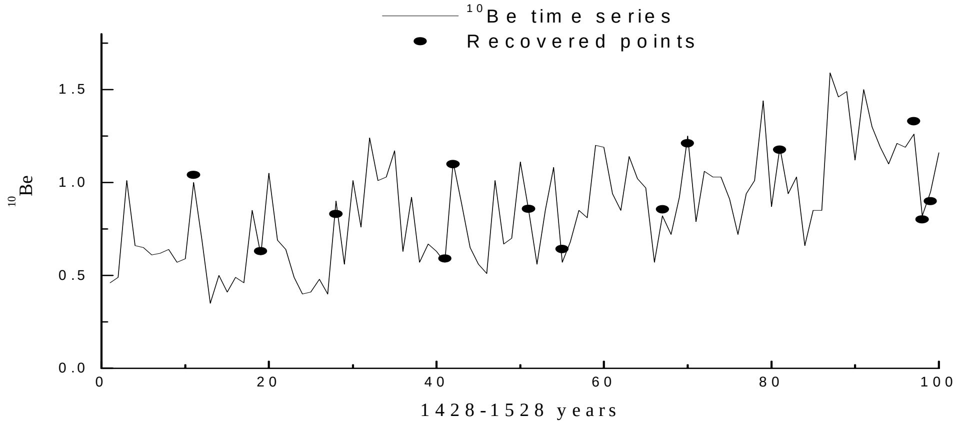

Figure 4 A fragment of a ^{10}Be time series, quasi-linear model, number of nodes: 6

Multi-Fractal Spectrum of ^{14}C Time Series

The multi-fractal spectrum *f*(α) is the characteristic customarily studied when dealing with multi-fractals. We calculate this quantity for the annual ^{14}C time series from 1510 AD–1954 AD (Stuiver and Braziunas 1993). In one part of the initial time series data values exist only for every second year (1891 AD–1910 AD), and there are some gaps (1911–1912, 1914, 1946). With the help of the method suggested above this missing data was recovered. Thus, we have obtained an equidistant time series that is applicable for multi-fractal analysis.

Let us note (Barreira et al. 1997), that the multi-fractal spectrum of singularities of strength α for the Borel finite measure μ on a compact set X is a function *f*(α) defined by a pair of functions (g,G). Here, g: X → [-∞, ∞] is a function which determines the level sets: $g : K^g_\alpha = \{x \in X : g(x) = \alpha\}$ and produces a multi-fractal decomposition.

$$X : X = \underset{-\infty\le\alpha\le+\infty}{Y} K^g_a$$

Let G be a real function, which is defined on $Z_j \subset X$ such as $G(Z_1) \leq G(Z_2)$, if $Z_1 \subset Z_2$. Then the multi-fractal spectrum is $f(\alpha) = G(K_\alpha^g)$. Let g be determined as point-wise dimension d_μ of measure μ at all points $x \in X$ for which the limit $g \equiv d_\mu(x) = \lim_{r \to 0} \left(\log \mu(B(x,r)) / \log r \right)$ exists, where $\mu(B(x,r))$ is a "mass" of measure in the ball of radius r centered at x. Since we have chosen $g = d_\mu$ we can drop the subscript g from further references to K_α^g.

Then $K_\alpha = \{x : d_\mu(x) = \alpha\}$, where the exponent α is the local density of μ. The singular distribution μ can then be characterized by the Hausdorff dimension of K_α, i.e. $f(\alpha) = G(K_\alpha) = \dim_H(K_\alpha)$. If μ is self-similar in some sense, $f(\alpha)$ is a well-behaved concave function (Falconer 1994).

To estimate $f(\alpha)$ we applied both the method of the partition sum and the method of direct calculation (Riedi 1997) on the ^{14}C time series. Figure 5 shows the $f(\alpha)$–spectrum. We see that the ^{14}C record has a large range of multi-fractal properties from 1.0 to 2.2.

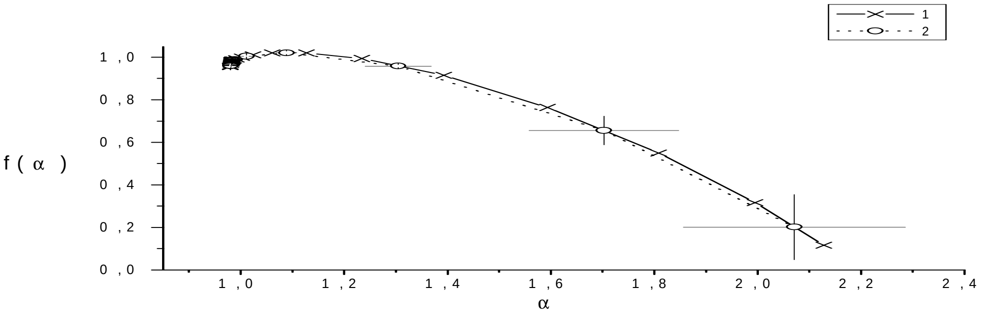

Figure 5 $f(\alpha)$-spectrum of annual ^{14}C time series calculated by two different methods: the partition sum (1) and direct calculation (2).

CONCLUSION

Our experiments have shown that the neural method for recovering of gaps in a time series is quite eligible for analysis of cosmogenic isotopes. This method allows to obtain equidistant time series, which can be analyzed by using the modern tools of non-linear analysis.

ACKNOWLEDGMENT

This research is supported by INTAS, grant nr 97–31008.

REFERENCES

Abarbanel HDI, Brown R, Sidorowich JJ, Tsimring L Sh. 1993a. The analysis of observed chaotic data in physical systems. *Review of Modern Physics* 65(4):1331–92.

Abarbanel HDI, Rabinovich MI, Sushchik MM. 1993b. Introduction to nonlinear dynamics for physicist. *World Scientific Lecture Notes in Physics* 53.

Anderson TW. 1971. *The statistical analysis of time series*. New York: John Wiley & Sons.

Barreira L, Pesin Ya, Schmeling J. 1997. On a general concept of multifractality: multifractal spectra for dimensions, entropies, and Lyapunov exponents. Multifractal rigidity. *Chaos* 7(1):27–38.

Beer J, Baumgartner St, Dittrich-Hannen B, Hauenstein J, Kubik P, Lukasczyk Ch, Mende W, Stellmacher R, Suter M. 1994. Solar variability traced by cosmogenic isotopes. In: Pap JM, Frohlich C, Hudson HS, Solanki SK, editors. *The sun as a variable star: solar and stellar irradiance variations.* Cambridge University Press. p 291–300.

Davis A, Marshak A, Wiscombe W. Cahalan R. 1994a. Multifractal characterizations of nonstationarity and intermittency in geophysical fields: observed, retrieved, or simulated. *Journal of Geophysical Re-*

https://doi.org/10.1017/S0033822200038224 Published online by Cambridge University Press

search 99(D4):8055–72.

Davis A, Marshak A, Wiscombe W. 1994b. Wavelet-based multifractal analysis of nonstationary and/or intermittent geophysical signals. In: Foufoula-Georgiou E, Kumar P, editors. *Wavelets in geophysics*. Academic Press. p 249–98.

Falconer KJ. 1994. The multifractal spectrum of statistically self-similar measures. *Journal of Theoretical Probability* 7(3):681–702.

Gorban AN, Makarov SV, Rossiev AA. 1998. Neural conveyor to recover gaps in tables and construct regression by small samplings with incomplete data. *Matematicheskoe Computernoe Obrazovanie* 5(II): 27–32. In Russian.

Gorban AN, Rossiev AA, Wunsch DC II. 2000. Self-organizing curves and neural modeling of data with gaps. *Proceeding of Neuroinformatics-2000*. Part 2. Moscow. p 40–6. In Russian.

Hastie T, Stuetzle W. 1988. Principal curves. *Journal of the American Statistical Association* 84(406):502–16.

Kramer MA. 1991. Non-linear principal component analysis using autoassociative neural networks. *AIChE Journal* 37(2):233–43.

Little RJA, Rubin DB. 1987. *Statistical analysis with missing data*. New York: John Wiley & Sons.

MacDonald GJ. 1989. Spectral analysis of time series generated by nonlinear processes. *Review of Geophysics* 27(4):449–69.

Rossiev AA. 1998. Modelling data by curves to recover the gaps in tables. *Nueroinformatics methods*. Krasnoyarsk: KGU Publishers. p 6–22.

Riedi RH. 1997. *An introduction to multifractals*. Rice University.

Sauer T, Yorke JA, Casdagli M. 1991. Embedology. *Journal of Statistical Physics* 65(3/4):579–616.

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

Stuiver M, Braziunas TF. 1993. Sun, ocean, climate and atmospheric $^{14}CO_2$, an evaluation of causal and spectral relationships. *The Holocene* 3:289–305.

https://doi.org/10.1017/S0033822200038224 Published online by Cambridge University Press