GEODEZJA I KARTOGRAFIA
t. LI, z. 4, s. 227–238 (2002)

Marek Przyborski
*Andrzej Stateczny**

The Naval University of Gdynia
(81-919 Gdynia, Śmidowicza Str. 69)

* Maritime University of Szczecin

Classification of the signals presenting different representations of the same process

In this paper there are presented results of applying the methods of the time series analysis to the problem of recognizing small boats. It has been showed that the acoustic signals of the boats can be classified by means of clustering algorithms.

INTRODUCTION

The classification of collected during diagnostic tests data gives us a better view on process of exploitation of the mechanism. Thus we can distinguish differences in dynamics of the whole system and obtain information about its shape. The problem of data classification has been intensively studied in the literature, detail discussion can be find in the following [1–7]. The common approach is based on computing a set of characteristic features for each time series, and then by comparing these values, the investigated objects are classified to different groups. The same situation we may find during investigating data from diagnostic measurements. One of the most difficult situations is when we have to deal with very similar signals, and we do not have any clues about the relations between signals and corresponding classes. An example of such situation can be meet when we have signals obtained from the objects of the same type.

In this paper we are going to identify objects, by splitting the collection of measures into different clusters. On this base we can compare two recordings made with time interval and check how the dynamics differs.

Our data contains different representations of the same process, they came from the hydro-acoustic experiment with different types of boats. From the variety of possible characteristics we have chosen the hydro-acoustic one, due to its attractiveness from the diagnostics point of view. During the sea trial we can record data, which presents the whole dynamics of the system (boat), tests at harbour cannot reveal such a results. Interaction between sea environment and the hull, including different mechanism on board, all this can give us better view on the changes in dynamics of the system. All those features are reflected in the hydro-acoustical signals.

1. *Experimental data*

The fingerprints of each ship are the unique vibrations of the hull, when the sea surface interacted with it while the ship is moving. If the distribution of those vibrations is totally random then hydro-acoustical signals should have also stochastic component and all the attempts to investigate this phenomenon by using deterministic methods must fail due to the stochastic nature of this process. By searching for these signatures we can answer the question about the vibration distribution and this would substantiated using nonlinear time series analysis methods to investigate hydro-acoustic phenomenon.

So far many attempts have been made to detect behavior characteristic for deterministic systems in the data coming from the real world. Finite number of points as well as finite resolution of those kind of data makes this investigation difficult to obtain satisfactory results. In this particular case we would like to use the method of surrogate data [8, 9], which provides statistical test for the null hypothesis that the data has been generated by a linear stochastic process. If this null hypothesis cannot be rejected, then most results of a nonlinear analysis are not correct. This kind of test is based on comparing the value of nonlinear measure for the original data and a number of surrogates.

2. *Description of the data*

In order to check the character of the vibration distribution we introduce hypothesis that the data was generated by a stationary Gaussian linear stochastic process. Computing nonlinear statistics on the data allows distinguishing deviations from the null hypothesis, our main purpose is to show that the original data x_n^o, differs significantly from the surrogates which are design as realizations of the null hypothesis.

The original data consists of samples of sound recorded on the test area while boat was approaching to the sensor. The sound made by the ship is detected by hydro-phones (sensor) which are connected to the standard PC computer equipped with the sound card. Recordings were downloaded into the hard-disk as a .wav files. The whole recording consists of 2.500.000 samples, therefore we split it, and for further considerations we have used only 30 segments of length 10000 points starting from the beginning of the recording. We decided to chose this part, due to its attractiveness from the analysis point of view. The boat is relatively far from the sensor and the signal is contaminated by the sound background of the sea. Both of them, boat and sea present a specific dynamics and in fact they can be considered as autonomous systems, however in the recorded signal they are represented by one observable.

The surrogate data is a set of data which mimic the original one, however consistent with the null hypothesis. For the presented conjecture we generated $\{x_n^k\}$, $k = 1,\ldots,B$ surrogates of the specific type, in our case they should be realization of the Gaussian linear stochastic process.

An example of the data is presents in the Fig. 1, this particular segment was recorded when the boat was in a distance of 1.5 nautical mile from the sensor, the lower panel presents an example of surrogate data created according to the scheme presented in the paper of Schreiber and Schmitz [10].

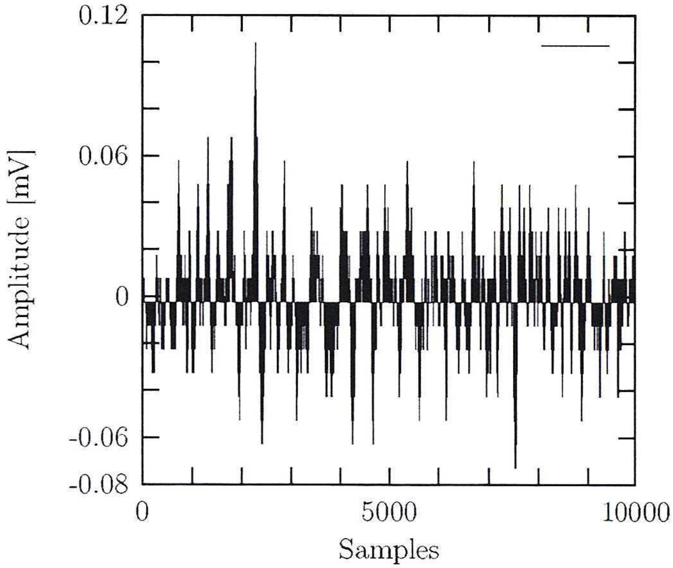

Fig. 1. Example of the data applied to the test

When creating tests for nonlinearity we should take into account two parameters. The first one, is its size α, which is the probability that the null hypothesis is rejected although it is in fact true. And the second, called level of significance $1 - p$ (usually the value of p is specify *a priori*, and then test is design with accordance to its value).

3. *Tests against determiinism*

3.1. Measures of nonlinearity

There exists several examples of different statistical methods (see for instance [8, 11–13]) which have been developed in order to reveal the nature of the considered time series.

We decided to apply two types of nonlinear statistics $t = t(\{x_n\})$:

1. As a first one, nonlinear prediction error with respect to a locally constant predictor F defined by

$$t^1(m, \tau, \varepsilon) = \left(\sum [x_{n+1} - F(x_n)]^2 \right)^{1/2} \tag{1}$$

The prediction is performed over one time step and it is done by averaging over the future values of all neighboring delay vectors closer than ε in m dimensions.

2. The second quantity it is cross-prediction errors, which can be expressed by the following formula

$$\sigma^2_{X,\,Y} = \frac{1}{L'} \sum_{k\,=\,(m\,-\,1)r\,+1}^{L\,-\,1} \| \vec{y}_{k+1} - Fx(\vec{y}_k) \|^2, \tag{2}$$

where: $L' = L - (m - 1)\tau - 1)$ – number of delay vectors, and F zeroth order model as it is proposed in the [11].

Calculating those nonlinear observable require using time delay embedding according to the following scheme, where embedding vectors in m dimensions are created by: $x_n = (x_{n\,-\,(m\,-\,1)\tau}, \ldots, x_n)$, – τ is the delay time.

4. *Results of the tests*

We generated the surrogate data according to the scheme described in [10], which is the appropriate method when the null hypothesis states that the data has been generated by a Gaussian linear stochastic process. This is the simplest description of a purely stochastic process, therefore we decided to applied this hypothesis.

The method is based on the phase randomized surrogate series $S = \{s_n, n = 1,\ldots,N\}$ which has the same power spectrum as the time series $X = \{x_n, n = 1,\ldots,N\}$. The temporal correlations in the original data are not preserved in the surrogates. The surrogate is obtained by determine the Fourier transform of the original data X, randomizing the phases, and inverting the transform.

The probability distributions of the nonlinear statistic $t = t(\{x_n\})$ remains unknown to us, therefore we applied a non-parametric, rank-based test, suggested in [14]. If the data deviates from the surrogates in a specific direction then we can reject the null hypothesis (one-side test). According to the size of the test α, we create $B = 1/\alpha - 1$ surrogate sets and then compute the test statistic t_o on the original data and on each of the surrogates (t_k, $k = 1,\ldots,B$).

For the prediction error, we expect nonlinearity in the data to appear in the lower values. Thus in this case we perform one-sided tests. All tests were carried out at the 95% level of significance, it means that for one-sided test we have created 19 surrogates.

Our data is a typical example of field measurement, it is strongly contaminated by the noise. In this particular case the nature of noise can have dual substance, the first source is a measurement noise, and the second one as we have already said is the natural sound background of the sea.

Assuming that the second type of noise and the distribution of unique vibration of the hull (when the hull is interacted with the sea surface) can be well described by the Gaussian linear stochastic process [15] we can expect that those features should be reflected in the results of the test. Simply in the presence of the stochastic components in our data test for nonlinearity should fail.

Conducted tests reveal that for one-sided test we achieved the 95% level of significance each time, results of one of the tests are presented in the Fig. 2.

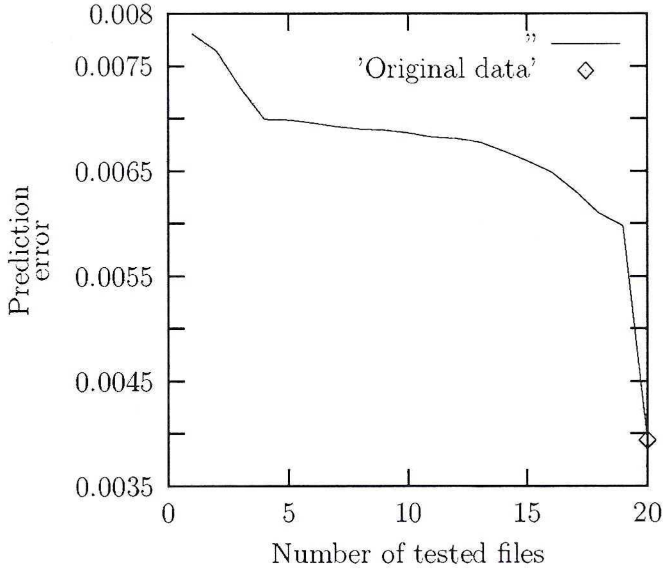

Fig. 2. Prediction error as a discriminating statistics

In order to precise our results we decided to apply method described in [5] which is based on the clustering algorithm. The main idea of using this algorithm in the test for nonlinearity, is to classify the total set of surrogates and original data, say K into two groups, from which one has only 1 element. As a dissimilarity measure cross-prediction errors were used. For conducting the test we used 9 surrogates and the original data, thus the probability that if the algorithm turned out the original data is $1/K = 0.1$, if it is true then we can reject the null hypothesis with the $(1-(1/K)) \times 100\% = 90\%$ of significance. For each combination of surrogates and original data from the set of 30 segments we obtained the rejection of the null hypothesis with the 90% level of confidence. An example of the answer given by the clustering algorithm is presented on the Fig. 3 and on the Fig. 4. As we have seen there are two clusters and one of them singled out the one element which contain the original data.

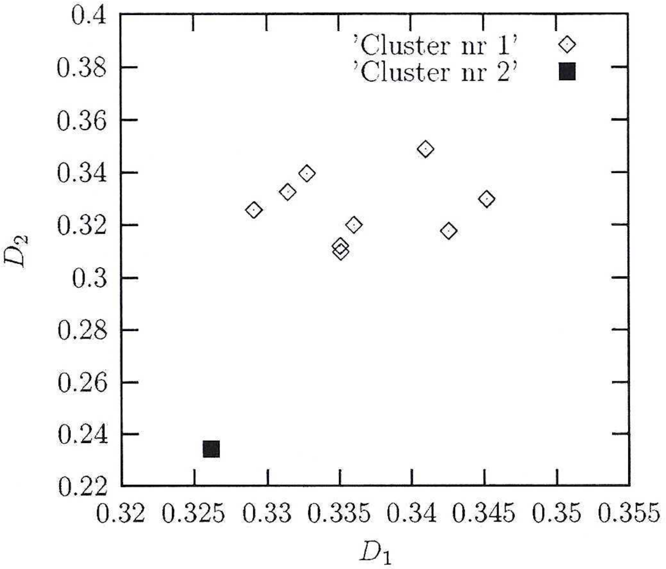

Fig. 3. Results of clustering – another file from the whole set of 30 recordings

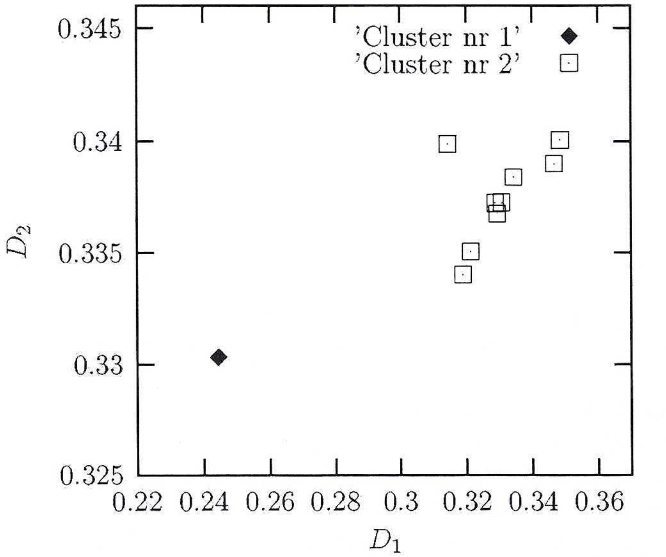

Fig. 4. Results of clustering – a sample file from the whole set of 30 recordings

We applied two nonlinear statistics in order to reveal the possible nature of considered signals. Taking into account results of conducted tests we can conclude that the null hypothesis can be rejected because original data can not be well described by the Gaussian linear stochastic process. It means that our assumption about the nature of hull's vibration is false, thus its distribution can be predicted due to deterministic nature of that process.

The null hypothesis which has been chosen in fact is the simplest one, however in our opinion it presents the main drawback in the field of investigation hydro-acoustics data. The possibility to reject the hypothesis of stochastic nature of that signals opens the way to extract the most important features of the object by means, for example nonlinear time series analysis. This allows us to characterize each type of object by the set of suitable parameters.

4.1. Classification of the data

Since the data represent the same phenomena, however each time recorded under slightly different conditions (type of the ship, differences in the shape of the propellers, differences in the hydrodynamical features of the hulls, meteorological conditions, sea level, different interactions between ships' hulls and sea surface) we assume that the attractors created by those systems (different classes of boats) are quite similar, therefore we are able to classify those signals to the corresponding class of similar dynamics.

We have eliminated the noise influence by the method of noise reduction with local projections [16–19], with the following parameters: embedding dimension $m = 20$, dimension of projection $q = 5$, size of the neighborhood $k = 30$ (for detail discussion of choosing those parameters see for example [16–19]. After filtration, our data is prepared for clustering, which is based on the following procedure:

- preprocessed data is used for computing similarity measure,
- then the results create the dissimilarity matrix,
- this matrix is used by the clustering algorithm to classify the data to the clusters.

In this particular case the data was prepared in the following mode: The whole collection of signals has been divided into parts of lengths 10000 points each. To compute the similarity measure those parts have been split into 40 sequences. We assumed that the number of signals which is used to calculate the dissimilarities, correspond to the number of clusters N, thus we do not have to chose the optimal partitioning into clusters.

If we want to classify K clusters into L classes first of all we have to define a membership index $\upsilon_{\mathrm{i}}^{\xi}$ vix (see the [5]) it is equal 1 when cluster i is located in class ξ and is equal 0 if otherwise, then the class definition can be described by

$$C^{(\xi)} = i : \upsilon_i^{\xi} = 1 \tag{3}$$

we want to form class where the average dissimilarity $d(x, y)$ of clusters within class is minimal, so if the number of clusters in class ξ is define as

$$|C^{(\xi)}| = \sum_{i=1}^{J} \upsilon_i^{(\xi)} \tag{4}$$

then in the class (ξ) the average dissimilarity of cluster *i* to other clusters in the same class is described by

$$D_i^{(\xi)} = \frac{1}{|C^{(\xi)}|} \sum_{j=1}^{J} \upsilon_j^{(\xi)} d\,(x,\, y) \tag{5}$$

thus, the average dissimilarity of clusters in class ξ can be expressed as

$$D^{(\xi)} = (1/|C^{(\xi)}|) \sum_{i=1}^{J} \upsilon_i^{(\xi)}\, D_i^{(\xi)} \tag{6}$$

of course for the total average we have to take the sum over all clusters, finally we can obtained the optimal partitioning into classes by finding the minimum of cost function which can be expressed as follow

$$E = LD = \sum_{\xi=1}^{K} \frac{1}{|C^{(\xi)}|} \sum_{i,j=1}^{J} \upsilon_i^{(\xi)}\, \upsilon_j^{(\xi)}\, d\,(x,\, y) \tag{7}$$

To compare the results obtained by the use of raw data and preprocessed by nonlinear noise reduction scheme, we show the scores obtained on clustering of 3 different signals. In the Fig. 5 results obtained on the raw data are present, the Fig. 6 shows the results with the clean data. The differences are very small. Thanks to our *a priori* knowledge about recordings we know that there should be 3 different signals.

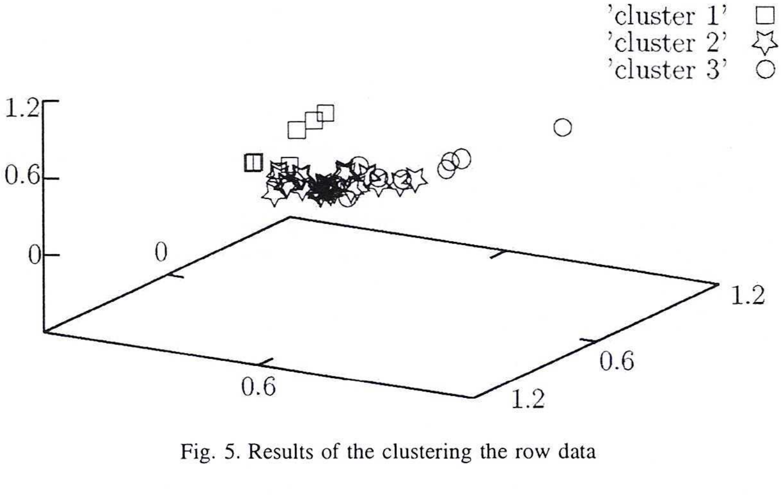

Fig. 5. Results of the clustering the row data

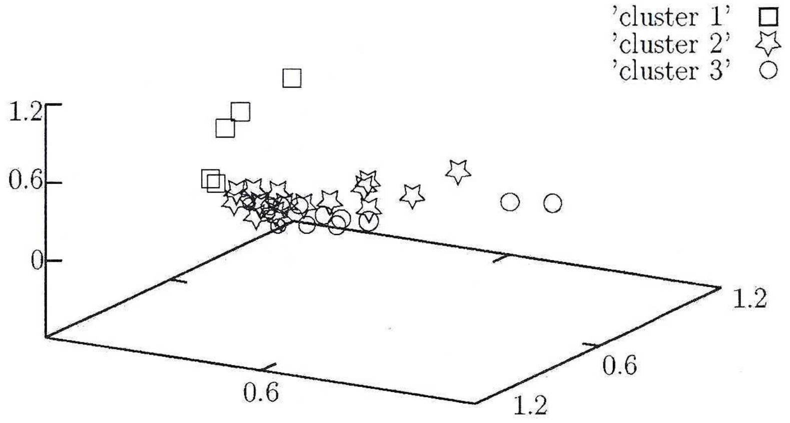

Fig. 6. Clustering the data after noise reduction

From Fig. 5 and Fig. 6 we cannot draw such a conclusion, therefore we decided to apply the singular value decomposition (SVD) method, in order to select the most important components from the recordings [2, 17, 18, 20, 21]. In our particular case we have chosen components. Conducted tests revealed that the most informative is the first principal component. Below there are results of clustering algorithm conducted on signals processed by SVD (see Fig. 7).

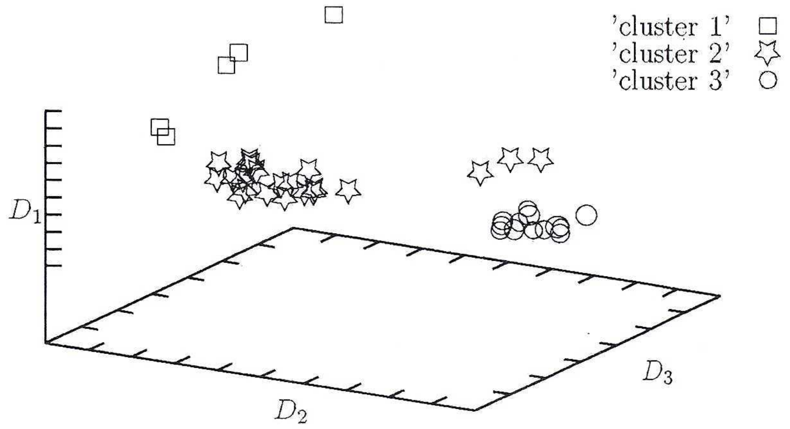

Fig. 7. Clustering the data after processing by SVD method

CONCLUSIONS

The above considerations lead us to the conclusion that classifying the experimental data on the base of similarities in the dynamics is possible, and in such a difficult case as hydro-acoustic signals which contain a very high level of noise applying nonlinear methods lead to satisfying results. However it is necessary to add that one of the most important problems remains still open – methods of noise reduction.

As we said the main application we see among the diagnostics methods. Characteristics obtained on machine working under different regime at sea (not during laboratory tests) are more reliable than tests in the harbour. We can diagnosis the power transmission system of the ship in its natural environment (at sea). Possibility to distinguish region with different dynamics gives us tool to analyze interaction between different devices on board. By means of cross-correlation integral we are able to check how differ certain parameters from the last test. Of course possibility to obtain hydrodynamics characteristics of the hull and propeller during the sea trial is also very important. This kind of analysis hydro-acoustics data can be very helpful in the field of naval architecture.

Acknowledgements

We wish to thank Rainer Hegger, Holger Kantz and Thomas Schreiber for using their package TISEAN which is available as a free software from its home page [48].

REFERENCES

[1] H. D. I. Abarbanel, *Analysis of observed chaotic data,* Springer, New York, 1996.

[2] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, L. Sh. Tsimring, *The analysis of observed chaotic data in physical systems,* Phys. Rev, A **38**, 3017, 1988.

[3] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, L. Sh. Tsimring, *The analysis of observed chaotic data in physical systems,* Rev. Mod. Phys., **65**, 1331, 1993.

[4] A. M. Albano, J. Muench, C. Schwant, A. I. Mees, P. E. Rapp, *Singular-value decomposition and the grassberger-procaccia algorithm,* Phys. Rev, A **38**, 3017, 1988.

[5] T. Buzug, G. Pfister, *Comparison of algorithms calculating optimal parameters for delay time coordinates,* Physica, D **58**, 127, 1992.

[6] M. Casdagli, S. Eubank, J. D. Farmer, J. Gibson, *State space reconstruction in the presence of noise,* Physica, D **51**, 52, 1991.

[7] C. Diks, J. C. van Houwelingen, F. Takens, J. DeGoede, *Reversibility as a criterion for discriminating time series,* Phys. Lett., A **201**, 221, 1995.

[8] C. Diks, W. R. van Zwet, F. Takens, J. DeGoede, *Detecting differences between delay vector distributions,* Phys. Rev., E **53**, 2169, 1996.

[9] J. P. Eckmann, D. Ruelle, *Ergodic theory of chaos and strange attractors,* Rev. Mod. Phys., **57**, 617, 1985.

[10] K. Fukunaga, *Introduction to Statistical Patern Recognition,* Academic Press, New York, 1990.

[11] P. Grassberger, R. Hegger, H. Kantz, C. Schaffrath, T. Schreiber, *On noise reduction methods for chaotic data,* CHAOS, **3**, 127, 1993.

[12] P. Grassberger, T. Schreiber, C. Schaffrath, *Nonlinear time sequence analysis,* Int. J. Bifurcation and Chaos, **1**, 521, 1991.

[13] R. Hegger, H. Kantz, T. Schreiber, *Practical implementation of nonlinear time series methods: The tisean package,* CHAOS, **9**, 413, 1999.

[14] F. Takens in D. A. Rand, L.-S. Young eds. Dynamical systems, and turbulence. *Detecting strange attractors in turbulence, Lecture notes in mathematics,* Vol. 898, Springer, New York, 1981.

[15] L. Jaeger, H. Kantz, *Unbiased reconstruction underlying a noisy chaotic time series,* CHAOS, **6**, 440, 1996.

[16] I. T. Jolliffe, *Principal component analysis,* Springer, New York, 1986.

[17] J. Kadtke, *Classification of highly noisy signals using global dynamical models,* Phys. Lett., A **203**, 196, 1995.

[18] H. Kantz, *Quantifying the closeness of fractal measures,* Phys. Rev., E **49**, 5091, 1994.

[19] H. Kantz, T. Schreiber, *Nonlinear time series analysis,* Cambridge University Press, Oxford, 1997.

[20] H. Kantz, T. Schreiber, I. Hoffmann, T. Buzug, G. Pfister, L. G. Flepp, J. Simonet, R. Badii, E. Brun, *Nonlinear noise reduction: A case study on experimental data,* Phys. Rev., E **48**, 1529, 1993.

[21] D. Kaplan, L. Glass, *Understanding nonlinear dynamics,* Springer, New York, 1995.

[22] L. Kaufman, P. J. Rousseeuw, *Finding Groups in Data, an introduction to cluster analysis,* Wiley, New York, 1990.

[23] E. J. Kostelich, T. Schreiber, *Noise reduction in chaotic time series data: A survey of common methods,* Phys. Rev., E **48**, 1752, 1993.

[24] E. J. Kostelich, J. A. Yorke, *Noise reduction in dynamical systems,* Phys. Rev., A **38**, 1649, 1988.

[25] D. Kugiumtzis, *State space reconstruction parameters in the analysis of chaotic time series the role of the time window length,* Physica, D **96**, 13, 1996.

[26] D. Kugiumtzis, *Assessing different norms in nonlinear analysis of noisy time series,* Physica, D **105**, 62, 1997.

[27] W. Liebert, H. G. Schuster, *Proper choice of the time delays for the analysis of chaotic time series,* Phys. Lett., A **142**, 107, 1989.

[28] R. Moeckel, B. Murray, *Measuring the distance between time series* Physica, D **102**, 187, 1997.

[29] E. Ott, T. Sauer, J. A. Yorke, *Coping with chaos,* Wiley, New York, 1994.

[30] D. B. Percival, A. T. Walden, *Spectral Analysis For Physical Applications,* Cambridge University Press, Cambridge, 1993.

[31] M. B. Priestley, *Non-linear and non-stationary time series analysis,* Academic Press, London, 1988.

[32] P. E. Rapp, A. M. Albano, T. I. Schmah, L. A. Farwell, *Filtered noise can mimic low-dimensional chaotic attractors,* Phys. Rev., E **47**, 2289, 1993.

[33] T. Sauer, J. Yorke, *How many delay coordinates do you need?* Int. J. Bifurcation and Chaos, **3**, 737, 1993.

[34] T. Sauer, J. Yorke, M. Casdagli, *Embedology,* J. Stat. Phys., **65**, 579, 1991.

[35] T. Schreiber, *Detecting anad analysing non-stationarity in a time series using nonlinear cross predictions,* Phys. Rev. Lett., **78**, 843, 1997.

[36] T. Schreiber, *Constrained randomization of time series data,* Phys. Rev. Lett., **80**, 2105, 1998.

[37] T. Schreiber, *Interdisciplinary applications of nonlinear time series methods,* Phys. Reports, **1**, 1999.

[38] T. Schreiber, P. Grassberger, *A simple noise-reduction method for real data,* Phys. Lett., A **160**, 411, 1991.

[39] T. Schreiber, H. Kantz in Y. Kravtsov, J. Kadtke eds., *Observing and predicting chaotic signals: Is 2% noise too much? in Predictability of complex dynamical systems,* Springer, New York, 1996.

[40] T. Schreiber, H. Kantz, *Noise in chaotic data: Diagnosis and treatment,* CHAOS, **5**, 133, 1995.

[41] T. Schreiber, A. Schmitz, *Immproved surrogate data for nonlinearity tests,* Phys. Rev. Lett., **77**, 635, 1996.

[42] T. Schreiber, A. Schmitz, *Classification of time series data with nonlinear simmilarity measures,* Phys. Rev. Lett., **79**, 1475, 1997.

[43] T. Schreiber, A. Schmitz, *Discrimination power of measures for nonlinearity in a time series,* Phys. Rev., E **55**, 5443, 1997.

[44] J. Theiler, D. Prichard, *Generating surrogate data for time series with several simultaneously measured variables,* Phys. Rev. Lett., **73**, 951, 1994.

[45] J. Theiler, D. Prichard, *Constrained-realization monte-carlo method for hypothesis testing,* Physica, D **94**, 221, 1996.

[46] H. Tong, *Non-linear time series analysis,* Oxford University Press, Oxford, 1990.

[47] A. S. Weigend, N. A. Gershenfeld, *Time series prediction: Forecasting the future and understanding the past,* Santa Fe Institute Studies in the Science of Complexity, Proc. Vol. XV, Adsison-Wesley, Reading, MA, 1993.

[48] http://www.mpipks-dresden.mpg.de/tisean

Received June 11 2002
Accepted October 10, 2002

Marek Przyborski
Andrzej Stateczny

Klasyfikacja sygnałów zastosowanych w różnych procesach

Streszczenie

Artykuł prezentuje rezultaty zastosowania metod analizy szeregów czasowych do rozwiązywania problemu rozpoznawania małych łodzi.

Wykazano, że sygnał hydroakustyczny generowany przez łodzie może być klasyfikowany przy zastosowaniu algorytmu klasteryzacyjnego.

Марек Пшыборски
Анджей Статечны

Классификация сигналов представляющих разные представления того-же процесса

Резюме

В статье представлены результаты применения методов анализа временных рядов по проблеме распознания небольших лодок. Доказано, что акустические сигналы лодок могут быть классифицированы при помощи кластерных алгоритмов.