HU9316203

KFKI-1993-02/G
REPORT

G. S. SRINIVASAN
K. KRINIZS
G. PÓR

ANOMALY DETECTION IN OECD BENCHMARK DATA USING CO-VARIANCE METHODS

Hungarian Academy of Sciences
CENTRAL RESEARCH INSTITUTE FOR PHYSICS
BUDAPEST

KFKI-1993-02/G
REPORT

ANOMALY DETECTION IN OECD BENCHMARK DATA USING CO-VARIANCE METHODS

G. S. Srinivasan*, K. Krinizs**, G. Pór***

* a visiting scientist from IGCAR, Kalpakkam, 603 102, India
** CRIP - Atomic Energy Research, Budapest, H-1525, Hungary
*** Institute for Nuclear Techniques of TUS, H-1511, Hungary

HU ISSN 0368 5330

G. S. Srinivasan, K. Krinizs, G. Pór: Anomaly detection in OECD Benchmark data using CO-Variance Methods
KFKI-1993-02/G

ABSTRACT

OECD Benchmark data distributed for SMORN VI has been investigated for the purpose of anomaly detection. It is observed that statistical features extracted from covariance matrix of frequency components are very sensitive in terms of the level of anomaly detection.

АННОТАЦИЯ

Данные Bench-mark, полученные для сопоставления различных методов анализа шумов, в процессе подготовки к конференции SMORN-VI, были использованы для обнаружения аномалий. Исследования показали, что статичнские характеристики, вычесленные из ковариационной матрицы частотных составляющих, являются очень чувствительными к аномалиям и открывают возможность для установления уровня сигнализации.

KIVONAT

A SMORN VI konferencia előkészítésére kibocsátott Bench-mark adatokat használtuk anomália detektálás-vizsgálat céljára. Vizsgálataink megmutatták, hogy a frekvencia összetevőkből képzett kovariancia mátrix további statisztikus derivátumai igen érzékenyek az anomáliákra és jól definiált alarm szintek képzését teszik lehetővé.

Introduction

At the Specialists Meeting in Reactor Noise (SMORN V) held in Munich, the results of an artificially generated reactor noise benchmark analysis was presented (Hoogenboom, 1988). However, the analysis methods adopted by the various participants differed in speed, sensitivity, and reliability. Following this, another benchmark was created (Hoogenboom, 1990) in which the system parameters were specified, so that the participants can exclude largely deviating signals due to misinterpretation of the data or the questions posed. The participants could check their noise analysis techniques for anomaly detection and stationary noise signals. In the case of anomaly detection, the participants were asked to estimate the earliest time of the onset of the anomalies by any of their detection methods. Theoretical aspects like false alarm rate and missing alarm rate were also asked for. The first results of this analysis were presented at IMORN 22, and more substantial results were presented a SMORN VI held at Gatlinburg.

In this report, we present the results of investigations of anomaly detection on the SMORN VI benchmark data using some new statistical features developed at IGCAR Kalpakkam. The programs were also developed at IGCAR, Kalpakkam.

Data description

The artificial noise data has been simulated by the digital reactor core simulator, RESIDEL, which simulates reactor core dynamics with thermal feedback. Three output signals were selected for the benchmark test, namely, fuel temperature fluctuations, coolant temperature fluctuations and reactor power fluctuations. Initially, signals were created in the digital form, and from that analog signals created by a digital - to - analog converter after low pass filtering. More details about the simulator are given by Hoogenboom (1989).

The actual layout of the data is as described in Fig. A. The anomaly was introduced in one of the model parameters which affected all three signals. They are inter-related by complicated transfer functions. The signals last totally for 9067 seconds. There are two anomalies present. The first anomaly is in the form of a step, onsets at 4039 seconds and lasts for 307.2 seconds. The second anomaly is the form of a ramp, onsets at 5554.6 seconds and continuously increases for a period of 1699.8 seconds.

The data used for the present analysis has been sampled from the original analog tape with a sampling frequency of 50 Hz. at KFKI. Four portions of the time file are used for analysis. Fig.B gives the description of the time files. Each time file has 61440 (60 K) points in all the three noise signals, and this corresponds to 1228.8 seconds of data in real time.

It is here important to mention the relevance of sampling time in the present data with reference to the analysis methods. Data with a higher sampling rate gives a more accurate indication of the onset of the anomaly. Most of the participants of the Benchmark analysis used SPRT method for the anomaly detection, wherein apart from the sampling time, initial values of false alarm and missing alarm rates also matter. This was more so in the case of the second anomaly (Husemann, 1990).

For the purpose of analysis, various statistical features were employed, as will be discussed below. Analysis was performed on Fuel Temperature Noise signals and Reactor Power Noise signals. Anomaly detection was possible only for the Reactor Power Noise signals. Hence, only these results are presented in this report. Here again, considering the time limitation and enormous amount of computer requirement, only the onset time and the duration time have been reported, and calculations for spurious alarm rate and missing alarm rate have not been done. However, in the case of the second anomaly, it may be noted that the time file is shorter than the extent of the anomaly (the time file begins before the anomaly and ends before the end of the anomaly) and hence only time of onset has been reported.

Analysis

For the purpose of anomaly detection, the following statistical features were used:

- variance
- psd-sum
- determinant of covariance matrix
- trace of covariance matrix

Variance:

This is the normal variance of the data. The variance was estimated by averaging over 256 points in the time domain. This feature was applied first directly on the data. Fig. 1 shows the result of application during the first anomaly, reactor power noise, and we observe a good discrimination. Fig. 2 shows the application of this feature on the second anomaly, reactor power noise. Though we see an increasing trend, we cannot conclude from this that anomaly detection is possible. To extract better information, smoothening was performed on the variance data. In this case, a simple method of smoothening, in which averaging was done over ten points taken at a time, was sufficient. Fig. 3 shows the result of this smoothening, in which we can see a better representation of the anomaly. Fig. 4 and Fig. 5 show the application of variance directly, and after smoothening, on the stationary region. This confirms the fact that a similar trend does not exist during normal operation, and what we observed in Fig. 3 is indeed the representation of the anomaly.

Psd-sum:

This feature is the simple sum of sensitive components in the power spectrum. Comparison of the power spectrums between the normal and anomaly noise regions had displayed sensitivity in the region 2Hz. to 10 Hz. (Figs.C & D) The power spectra were estimated by averaging over 128 points in each block. The sum of the psd values from the 5th block to 26th block (1.56 Hz. to 10.16 Hz.) was chosen as the feature. Direct application of this feature showed good discrimination in the first anomaly (Fig. 6). However, in the case of the second anomaly (Fig. 7) we do not see any discrimination except a mild trend.

Determinant and Trace of Covariance matrix:

The construction of these features are explained in Srinivasan and Singh (1990), (1992) These features are evaluated in the following way:

- power spectrums are estimated throughout the time-series (in this case, they are estimated by taking 128 time domain points in each block)

- sensitive frequency components taken from each power spectrum (in this case, all 22 components in the range 1.56 Hz. to 10.16 Hz. were chosen)

- data matrix formed for covariance estimation by taking a sequence of power spectrums (in this case, 50)

- the covariance matrix evaluated

- determinant and trace of this matrix evaluated

The determinant and Trace are evaluated in a sequence by taking power spectrums on a sliding basis, and evaluating the successive covariance matrices.

The considerations of choosing the number of time domain points in each psd, the number of sensitive frequency components in each psd, and the number of psds. in the data matrix for covariance averaging, are as follows:

- there should be enough number of sensitive points in each psd. for getting a larger covariance matrix in tune with the computer's capability of evaluating the determinants within its numerical limits

- enough number of psds should be considered for each data matrix in tune with the averaging requirements for the evaluation of covariance

- the above two requirements should together satisfy the condition, that there should be enough number of power spectrums in the anomaly region for estimation of onset and duration time

Based on the above requirements 128 points were chosen in each power spectrum, (this results in 120 psds in the first anomaly region), all 22 sensitive points chosen in the sensitive frequency range, and 50 power spectrums chosen in each data matrix for covariance matrix evaluation. But as will be seen below, even this was no sufficient.

At first, this feature was applied directly for detection of the anomalies. Fig. 8 shows the application of determinant on the first anomaly and Fig. 9 that of Trace. Fig. 10 shows the application of determinant on the second anomaly and Fig. 11 that of Trace. We do not see any indication of the first anomaly both for determinant and trace, while in the case of the second anomaly, we see an increasing trend resembling the second anomaly.

The above mentioned features have proved to be very powerful on application to sodium boiling and sodium water leak noise data. (Srinivasan and Singh (1990), (1992)). In spite of optimizing the various numbers based on earlier mentioned considerations, these features have failed in the first anomaly region. The reason for non indication in the case of first anomaly is attributed as follows. The present data is sensitive in the low frequency region. For the purpose of covariance averaging over the power spectrums, the anomaly region has to be long enough to have enough amount of data for averaging, as explained earlier. The duration of the first anomaly is not large enough whereas the second anomaly is large enough. Hence we see a trend in the case of second anomaly and no indication in the case of first anomaly. Here again, the considerations of sampling time come in the picture. Higher sampling to enlarge the data set would only result in in the representation of the sensitive region and the low frequency range by a fewer number of points. This means a lower resolution in the sensitive frequency range. Hence, increasing the number of power spectrums by increasing the sampling rate is not the solution. To generate more number of power spectrums, two methods were employed.

In the first method, a linear interpolation is carried out between adjacent power spectrums. Of course, this method has some basic contradictions considering the stochastic nature of the signals. However, considering the advantage of big reduction in computing time for creating the extra power spectrums, this method was tried. Nine extra spectrums are created between the adjacent spectrums. The covariance matrices, determinant and trace are estimated as earlier. The result of this investigation is shown in Fig. 12 for the behaviour of the determinant during the first anomaly and Fig. 13 for trace during the first anomaly. The scale of the Y-Axis in Fig.1 is linear, but the values are logarithm of the original Determinant values. We observe a very good discrimination, particularly considering the fact that the values are actually logarithmic in nature. From Fig. 13, we observe that Trace also gives a high amount of discrimination in the case of the first anomaly. However, in the case of the second anomaly, only a trend of the anomaly is seen. Smoothening of the data may indicate more clearly the trend of the anomaly.

In the second method, power spectrums were estimated by sliding over the raw data values, and the covariance matrices are estimated in turn, by sliding over the power spectrums. The sliding was actually done by skipping 10 points in the raw data and taking psd. over the 128 points. Then we get a total of 6144 spectrums in the first anomaly file, instead of getting 480 spectrums in the direct sequential estimation. The results of this application are shown in Fig. 16 to Fig. 19. Here, the values for determinant are given in log. along the Y-axis. In the case of first anomaly, both determinant and trace give good discrimination. In the case of the second anomaly, we do not see a good discrimination.

On comparison of the data enlarging techniques, we see that there is a better representation of the anomalies in the data sliding technique than data interpolation technique. It should be noted that both have failed to indicate anything in the second anomaly region. The data sliding technique is more acceptable than the former since linear interpolations are not justified in stochastic signals. Of course, in the later case also, there are problems due to windowing But the later is more acceptable than the former.

Further discussions:

Table 1 gives the results of the time of anomaly onset and duration in the case of first anomaly, and onset time in the case of the second anomaly.

The following significant aspects need a deeper consideration:

- that the determinant gives an extremely high amount of discrimination in the anomaly noise signal, followed by trace

This is essentially because of the inter correlations between the frequency components which change when anomaly region is encountered. Apart from this, there is also a change in the individual values of the components with time. Both this result in the high discrimination when determinant is evaluated. As per trace, only the second reason holds, since this is simply the sum of the individual variance values of the frequency components. Hence the extent of discrimination is less in the case of trace.

- that in the first attempt of evaluating determinant and trace, whereas there is a failure in the first anomaly region, there is indeed some success in the second anomaly region

The reason for failure in the first anomaly region is already explained. In the second anomaly region, the extent of anomaly is large and hence there is less problem of data availability. Hence, we see a trend of anomaly in this case.

- that when data enlargement is done in both ways, there is success in the first anomaly region, whereas there is failure in second anomaly region.

This brings to light the fact that data enlargement infect only corresponds to less amount of time domain data for each evaluation of determinant and trace, and works well only when the anomaly is in the form of a step or impulse. In case of the second anomaly, since it is only a continuous rise, this does not work, whereas the first attempt has worked.

- the onset time and duration time in case of determinant and trace in the first anor egion, and the onset time in the second anomaly region have some offset from the actual values

This is due to covariance averaging. Since 50 spectrums are considered (in the first attempt) the uncertainty of evaluation would be time corresponding to 25 spectrums (64 seconds !). However when the data is enlarged, the uncertainty reduces considerably for obvious reasons.

Conclusions:

We conclude from the above analysis that the Determinant and Trace extracted out of covariance matrix from the sensitive components of power spectrum are very sensitive to anomaly noise signals and could be used for on-line applications. It is important to note that, to avoid uncertainty in declaration of the onset of anomaly, the determinant and trace should be evaluated from closely evaluated power spectrums (like in the sliding data case). However, when the noise signals are sensitive in the high frequency region, (in kHz. range), direct application of the above features could be effective in terms of level of detection and less uncertainty of declaration of onset time. (Srinivasan and Singh, (1990),(1992))

Acknowledgement

This work has been done in the frame of the Scientific and Technological Co-operation between the Indian and Hungarian Academies of Sciences. The activities were financed partly by Hungarian grant OMFB No 00559.

The authors thank Mrs. O. Glodi for her assistance in preparation of this report.

References:

1. Hoogenboom et al (1988) - Progress in Nuclear Energy vol. 21, pp: 815-824

2. Hoogenboom (1989) - Benchmark Test Description

3. G.S.Srinivasan et. al. (1990) - New statistical features sensitive to sodium boiling noise - Annals of Nuclear Energy Vol: 17, no: 3, pp: 135-138, 1990.

4. G.S.Srinivasan et. al. (1992) - New statistical features for sodium boiling and leak noise detection in liquid metal fast breeder reactors under poor signal to noise ratio conditions - to be published in annals of nuclear energy.

5. A.Husemann et al.(1991) "Application of Auto-Regressive program package on Benchmark data" - Summary report on deputation to KFKI Jul. 7 - Sept. 6, 1991.

Table - 1

Feature	First Anomaly Onset Time	First Anomaly Duration	Second Anomaly Onset Time	Second Anomaly Duration
Actual data	4039	307.2	5554.6	1699.8
Variance	4037.12	308.5	5553.42 (after smoothening)	-
Psd-sum	4037.0	307.8	-	-
Determinant				
-Direct	-	-	5738.8	-
-Interpolated	4038.92	366.39	-	-
-Sliding psd.	4026.8	325.8	-	-
Trace				
-Direct	-	-	5735.4	-
-Interpolated	4034.71	357.4	-	-
-Sliding psd.	4026.81	311.75	-	-

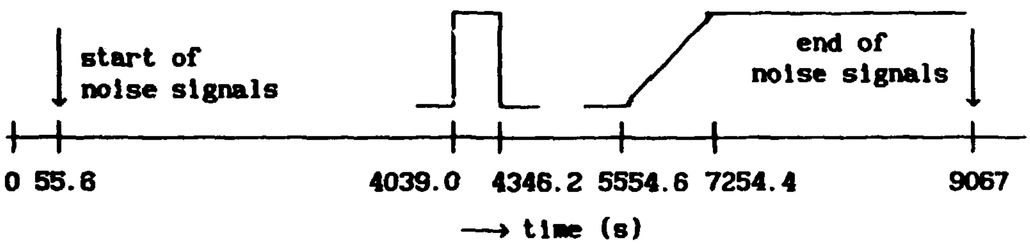
start of
noise signals
end of
noise signals
0 55.6
4039.0
4346.2
5554.6
7254.4
9067
⟶ time (s)

FIG. A

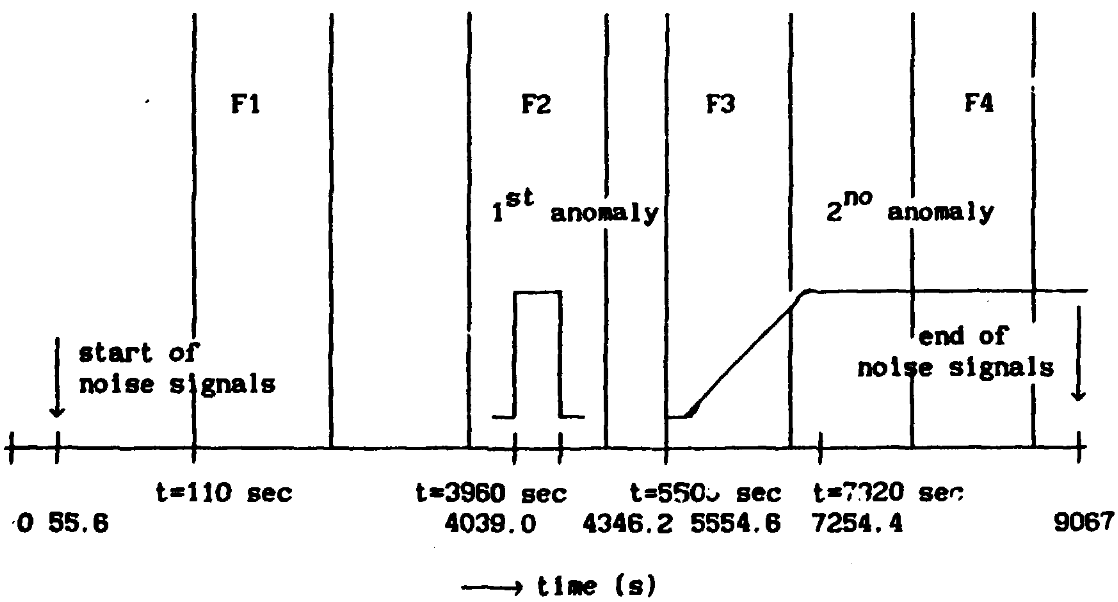
F1
F2
F3
F4
1st anomaly
2nd anomaly
start of
noise signals
end of
noise signals
t=110 sec
t=3960 sec
t=7320 sec
0 55.6
4039.0
4346.2 5554.6
7254.4
9067
⟶ time (s)

FIG. B

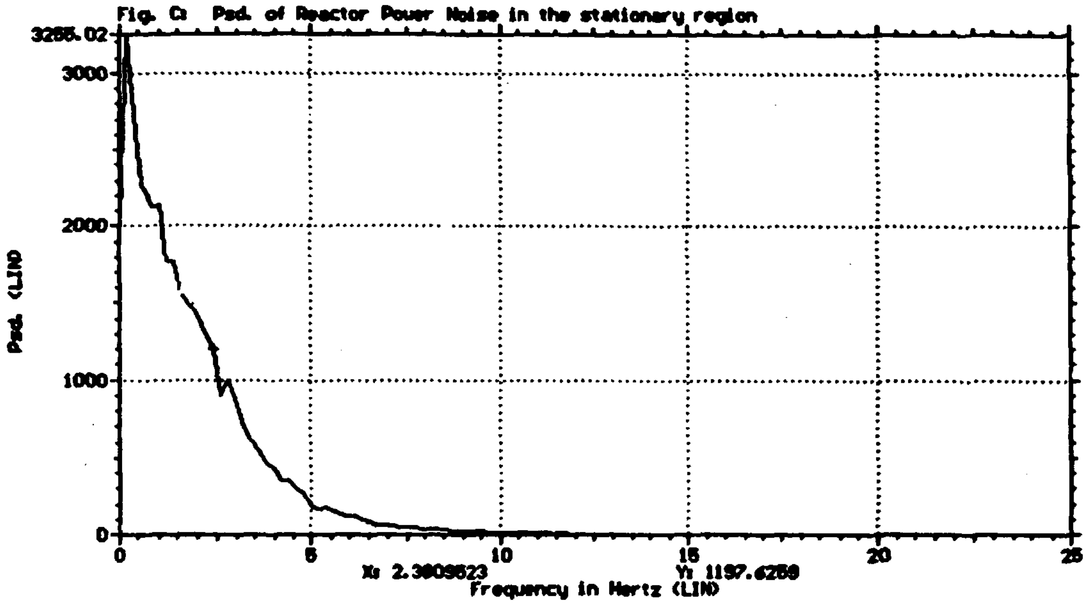

Fig. C Psd of Reactor Power Noise in the Stationary Region

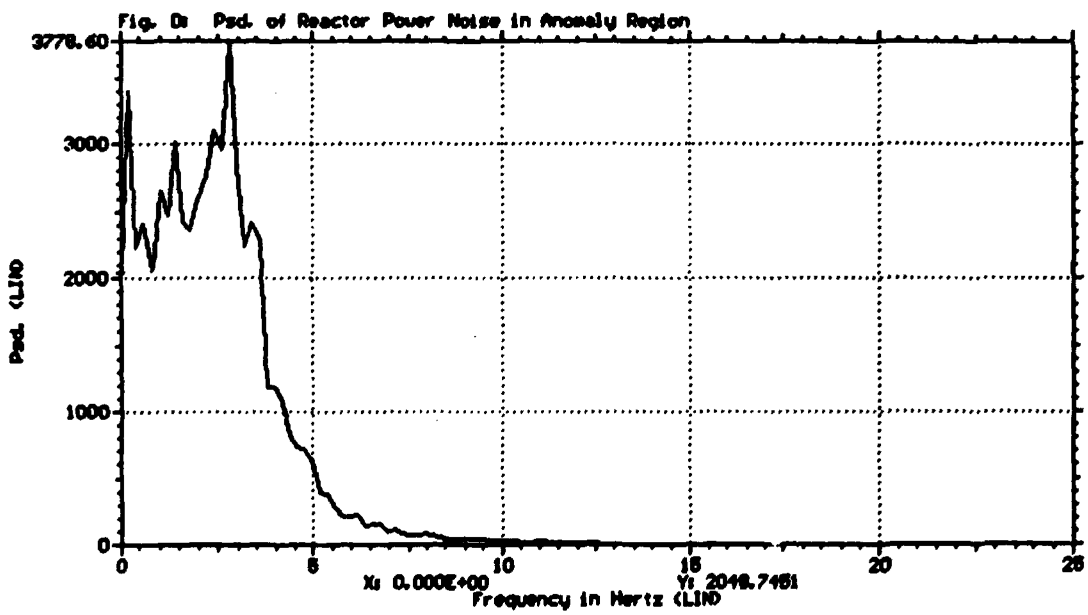

Fig. D Psd of Reactor Power Noise in the Anomaly Region

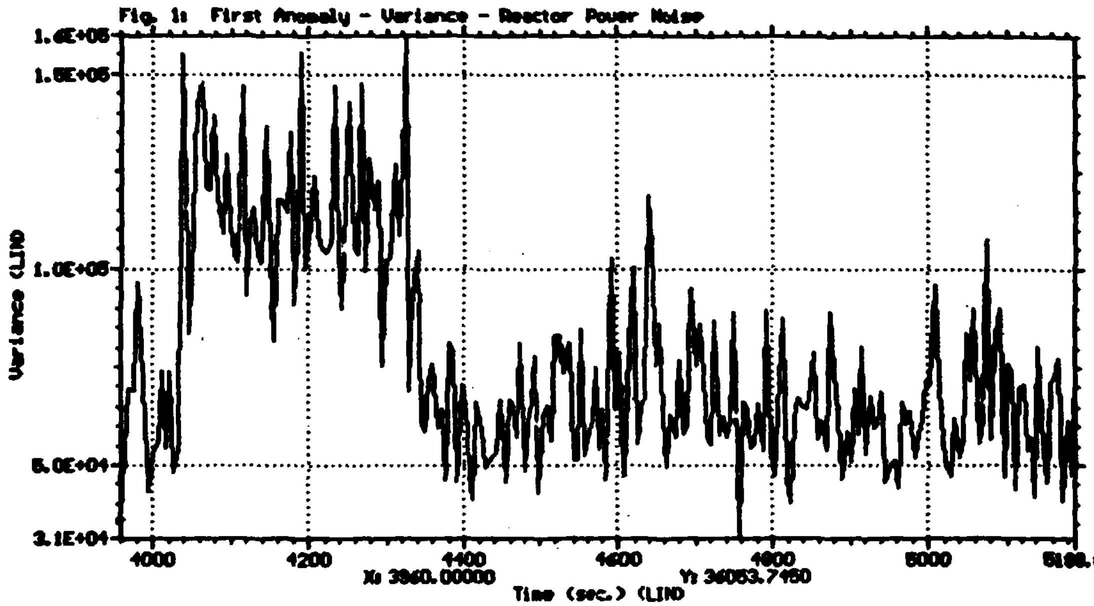

Fig.1. First Anomaly - Variance - Reactor Power Noise

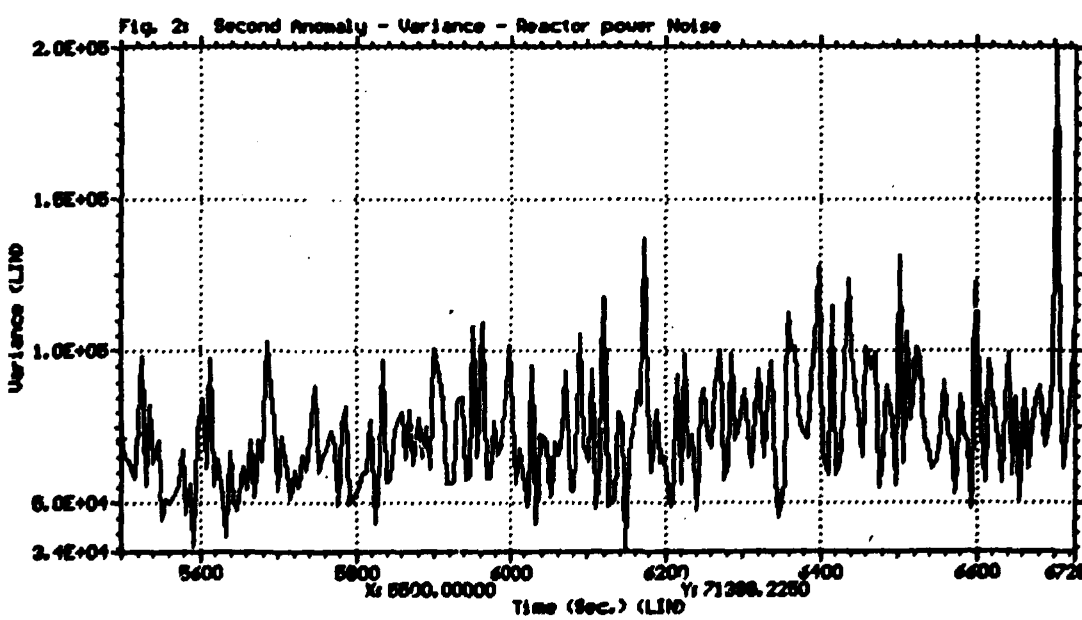

Fig.2. Second Anomaly - Variance - Reactor Power Noise

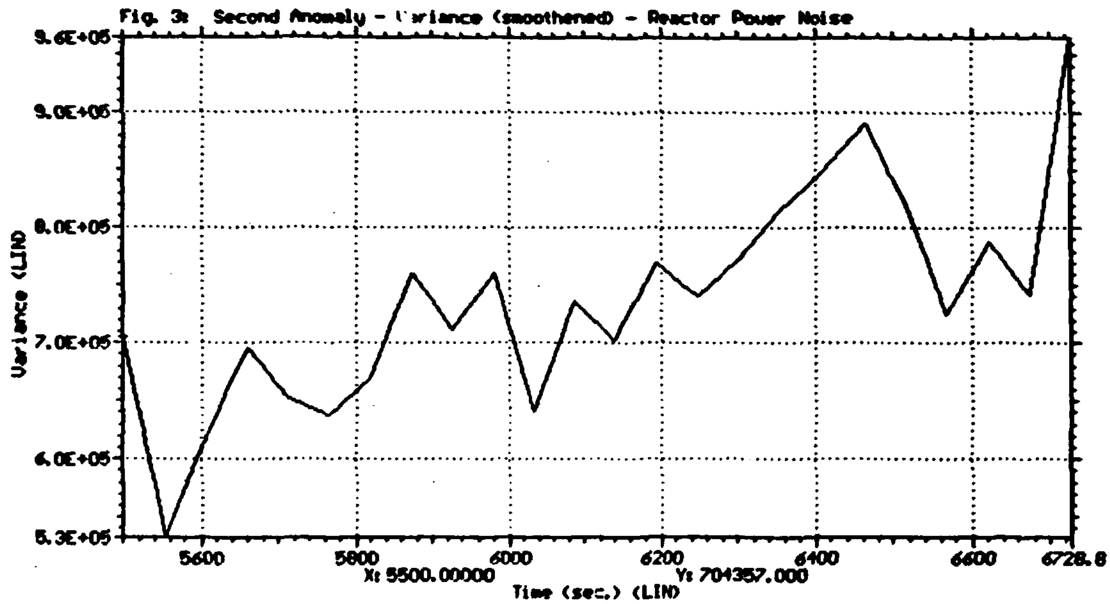

Fig.3. Second Anomaly - Variance (smoothened) - Reactor Power Noise

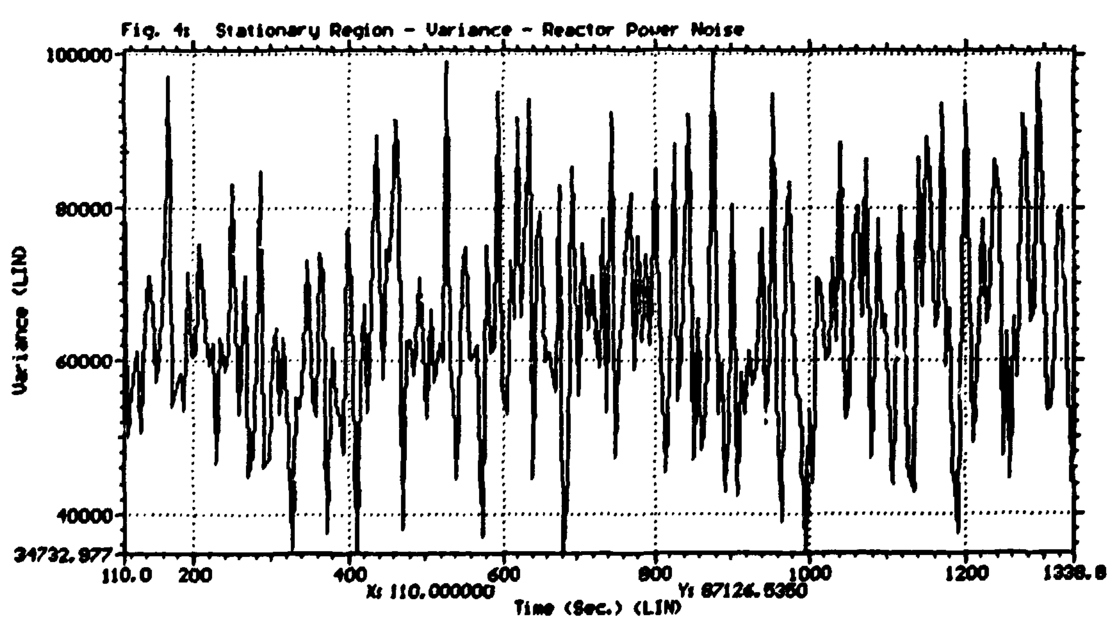

Fig.4. Stationary Region - Variance - Reactor Power Noise

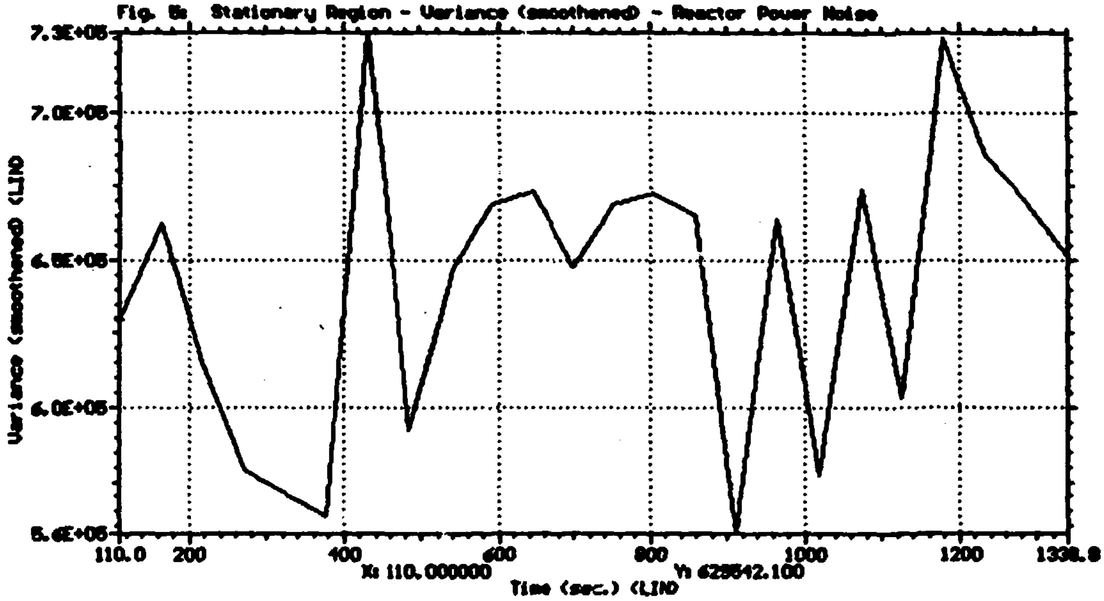

Fig.5. Stationary Region - Variance (smoothened) - Reactor Power Noise

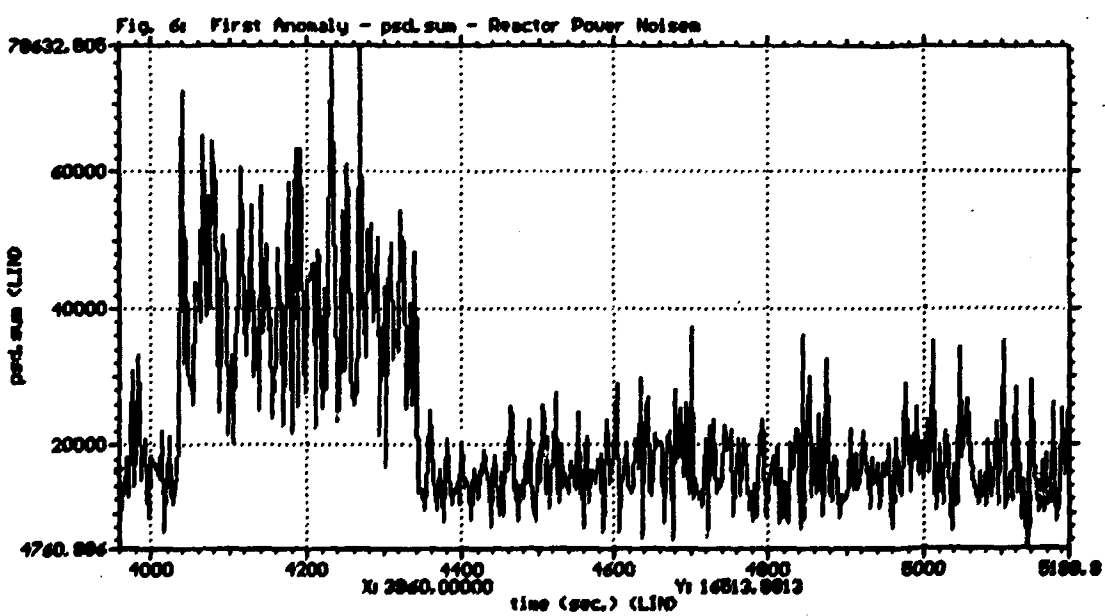

Fig.6. First Anomaly - psd. sum - Reactor Power Noise

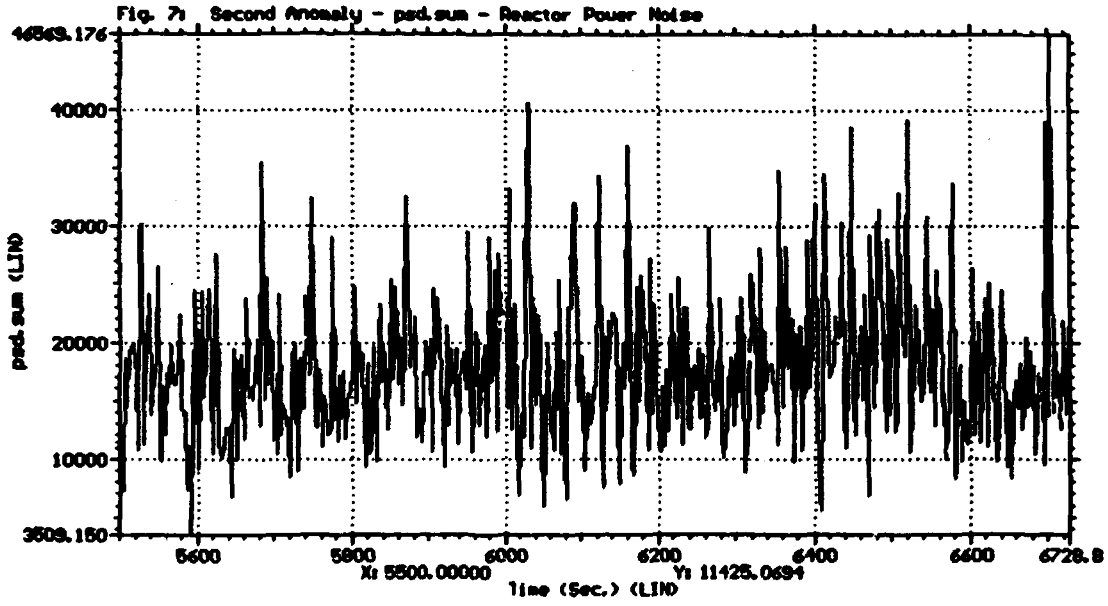

Fig.7. Second Anomaly - psd. sum - Reactor Power Noise

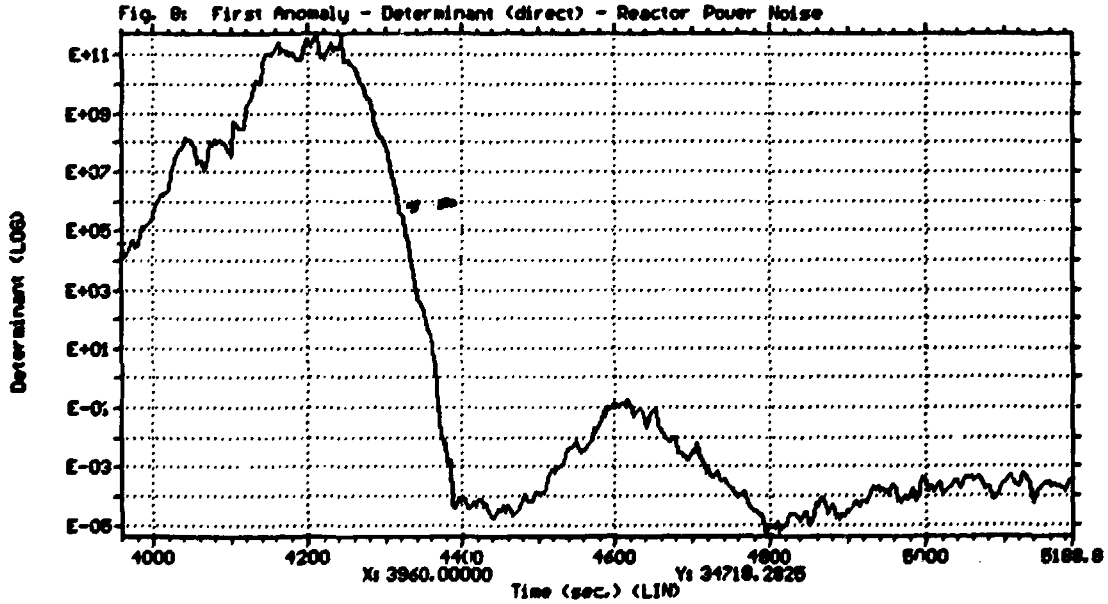

Fig.8. First Anomaly - Determinant (direct) - Reactor Power Noise

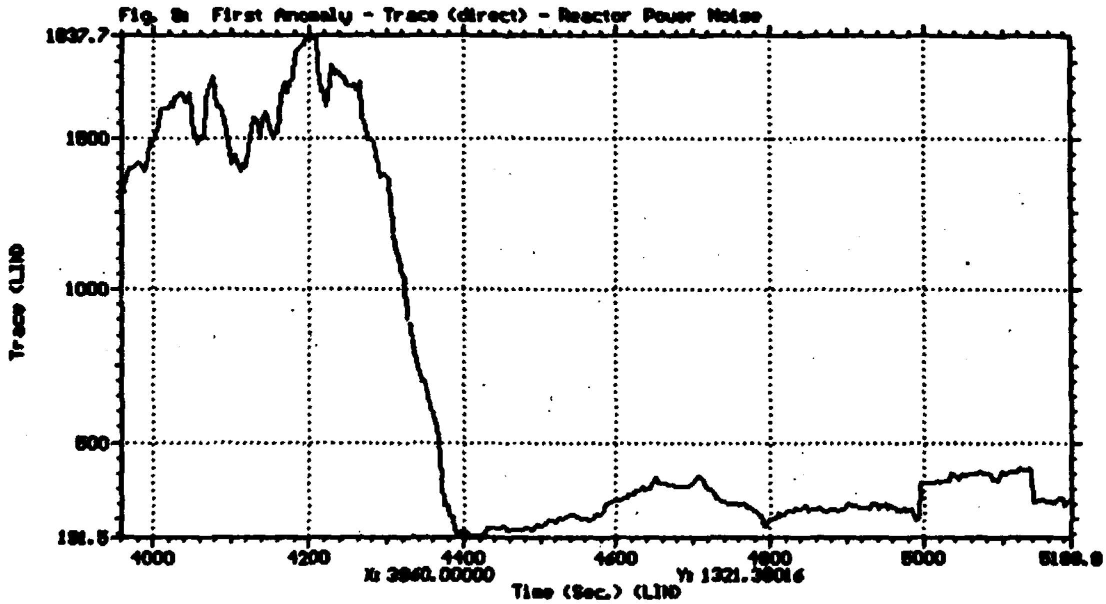

Fig.9. First Anomaly - Trace (direct) - Reactor Power Noise

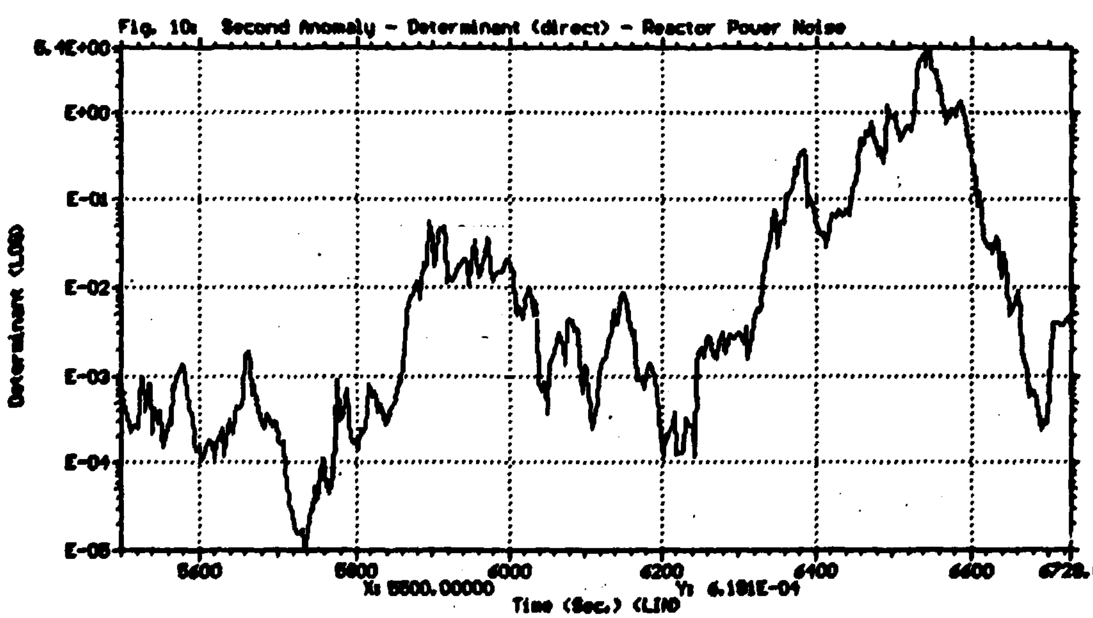

Fig.10. Second Anomaly - Determinant (direct) - Reactor Power Noise

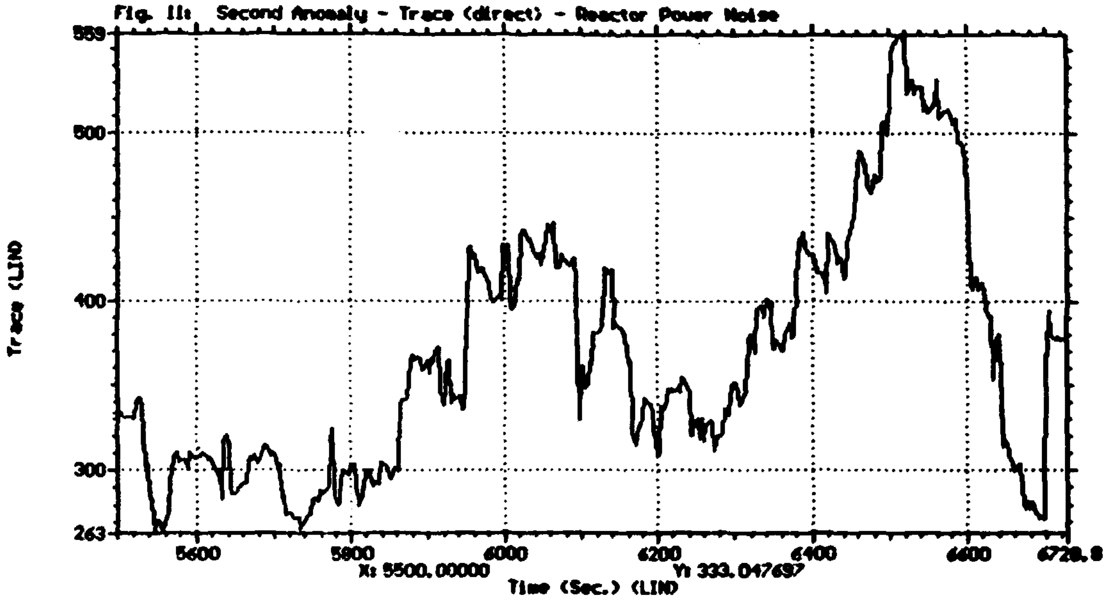

Fig.11. Second Anomaly - Trace (direct) - Reactor Power Noise

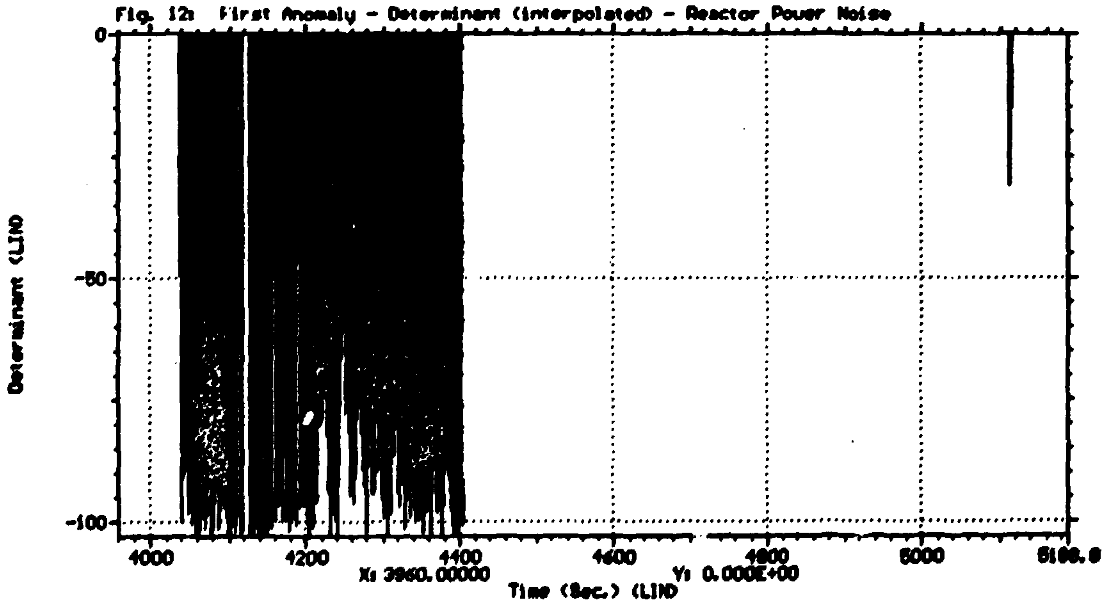

Fig.12. First Anomaly - Determinant (interpolated) - Reactor Power Noise

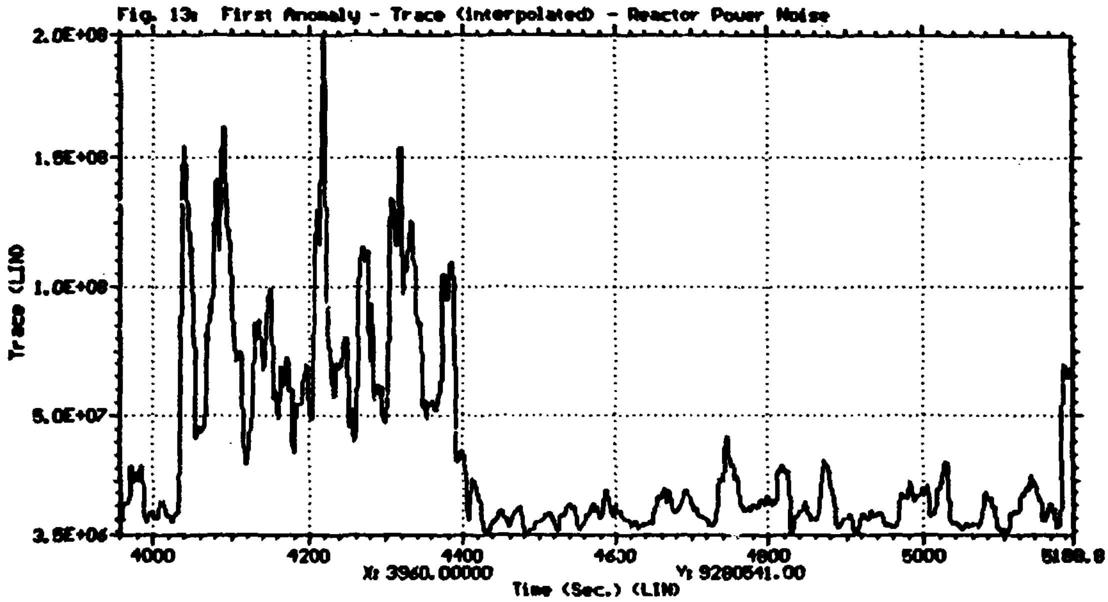

Fig.13. First Anomaly - Trace (interpolated) - Reactor Power Noise

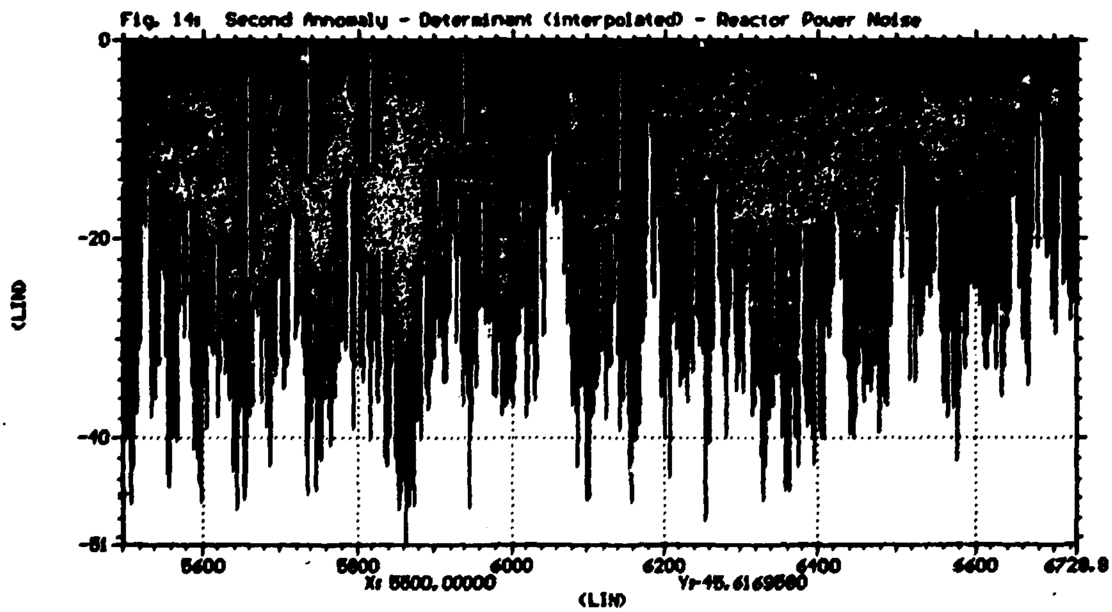

Fig.14. Second Anomaly - Determinant (interpolated) - Reactor Power Noise

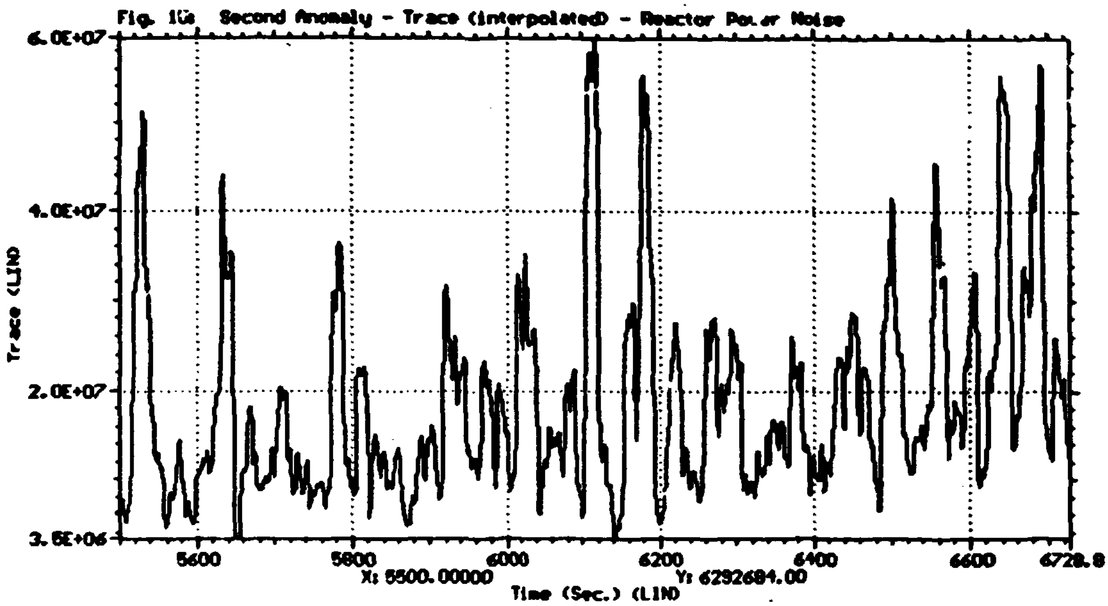

Fig.15. Second Anomaly - Trace (interpolated) - Reactor Power Noise

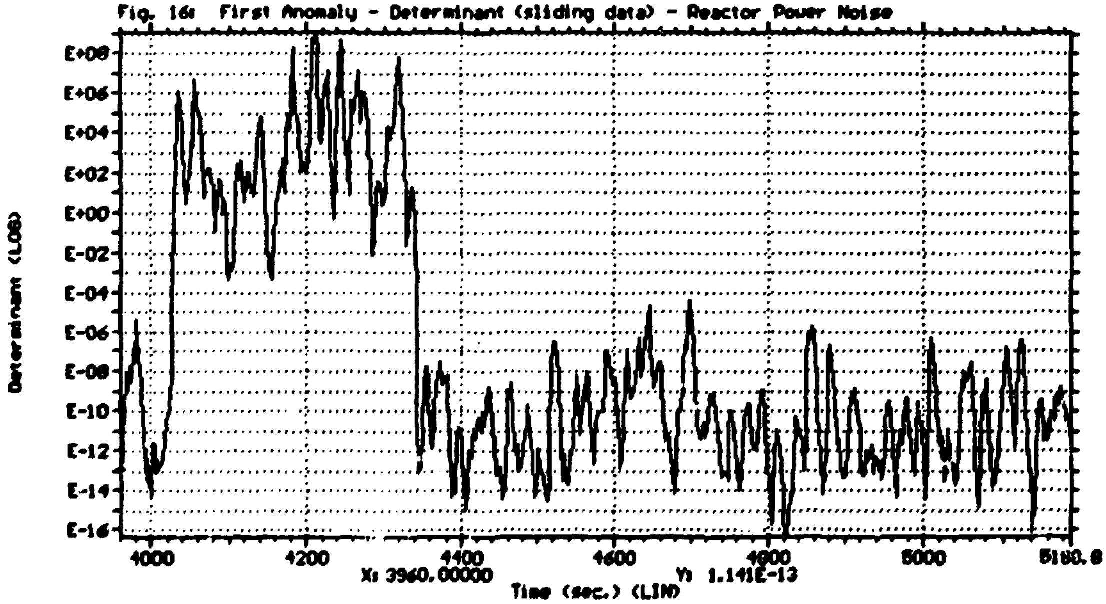

Fig.16. First Anomaly - Determinant (sliding data) - Reactor Power Noise

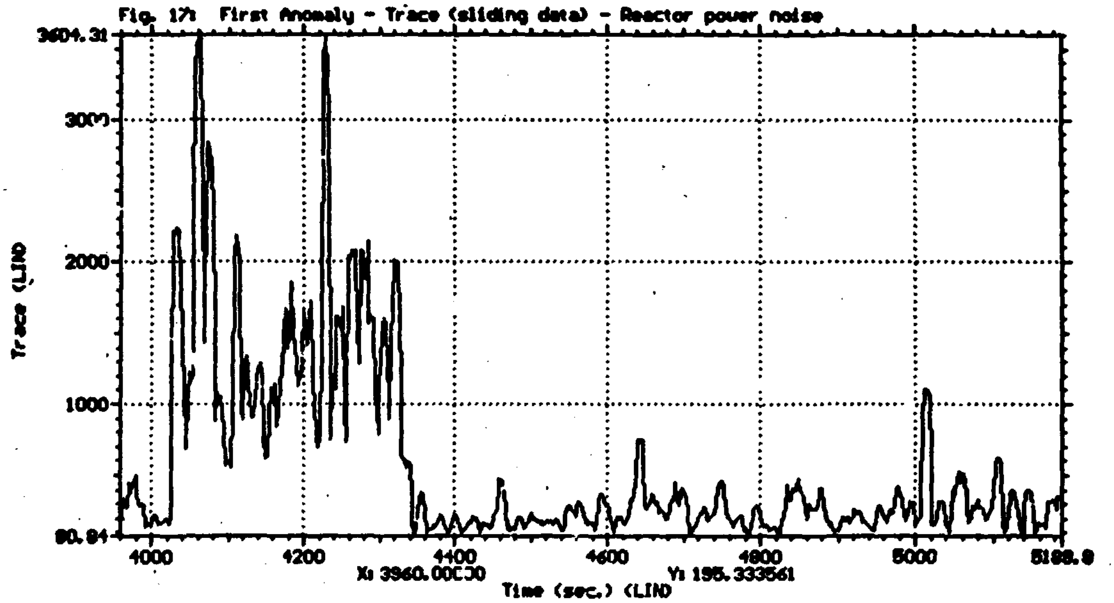

Fig.17. First Anomaly - Trace (sliding data) - Reactor Power Noise

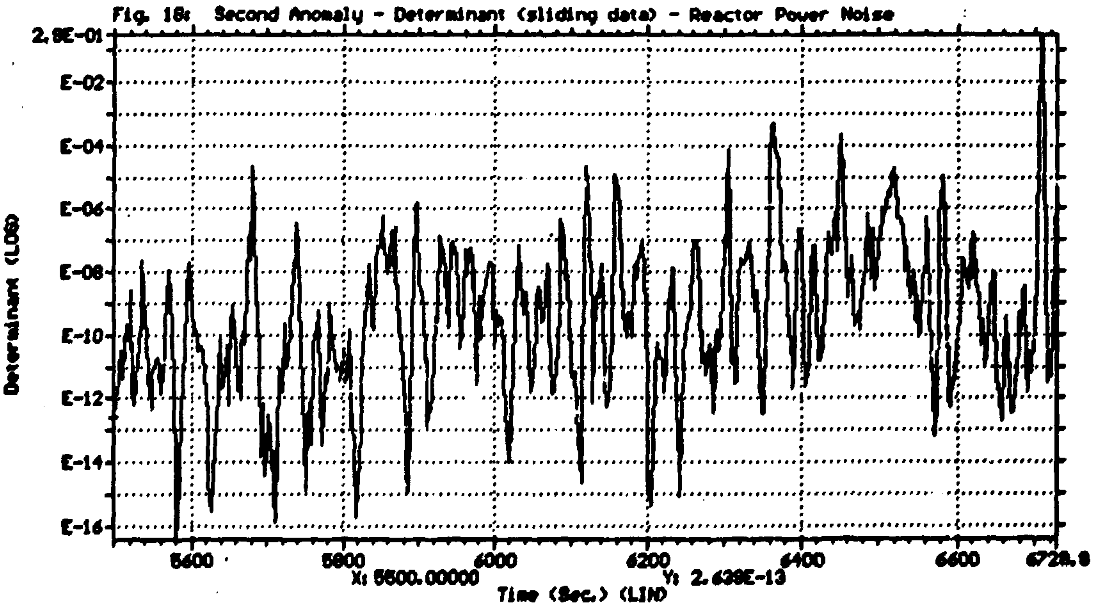

Fig.18. Second Anomaly - Determinant (slidi

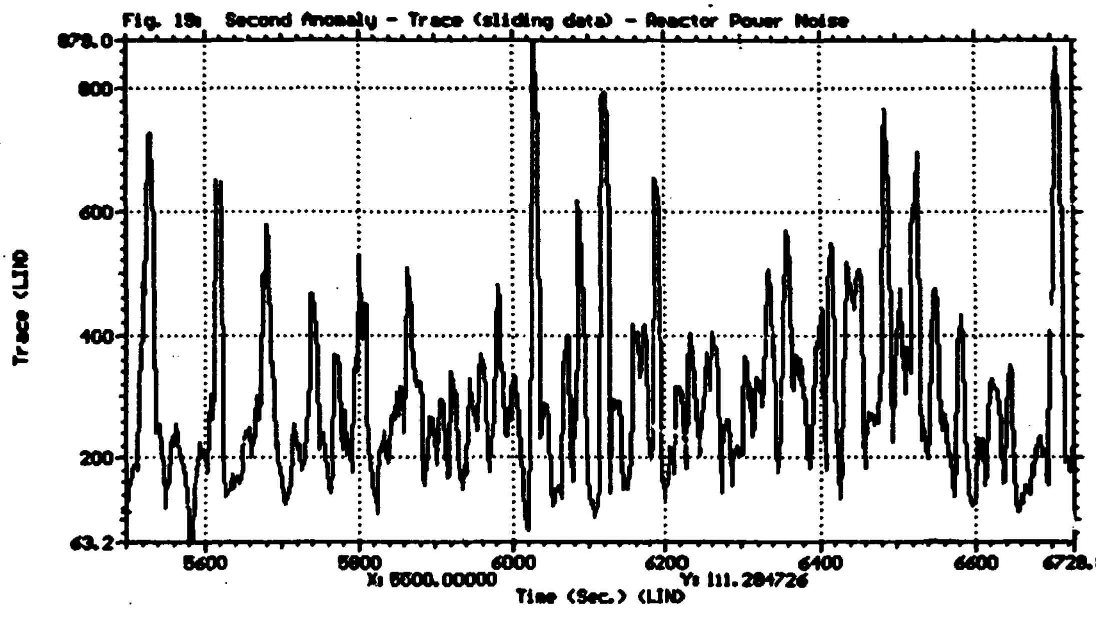

Fig.19. Second Anomaly - Trace (sliding data) - Reactor Power Noise

The issues of the KFKI preprint/report series are classified as follows:

A. Particle and Nuclear Physics
B. General Relativity and Gravitation
C. Cosmic Rays and Space Research
D. Fusion and Plasma Physics
E. Solid State Physics
F. Semiconductor and Bubble Memory Physics and Technology
G. Nuclear Reactor Physics and Technology
H. Laboratory, Biomedical and Nuclear Reactor Electronics
I. Mechanical, Precision Mechanical and Nuclear Engineering
J. Analytical and Physical Chemistry
K. Health Physics
L. Vibration Analysis, CAD, CAM
M. Hardware and Software Development, Computer Applications, Programming
N. Computer Design, CAMAC, Computer Controlled Measurements

The complete series or issues discussing one or more of the subjects can be ordered; institutions are kindly requested to contact the KFKI Library, individuals the authors.

Title and classification of the issues published this year:

KFKI-1993-01/M M. Tőrő, Gy. Csopaki, L. Zs. Varga, K. Tarnay & al. (eds): Case studies for protocol specification with proconsul

KFKI-1993-02/G G. S. Srinivasan, K. Krinizs, G. Pór: Anomaly detection in OECD Benchmark data using CO-Variance Methods

Kiadja a Központi Fizikai Kutató Intézet
Felelős kiadó: Gadó János
Szakmai lektor: Lux Iván
Nyelvi lektor: Lux Iván
Példányszám: 273
Készült a PROSPERITÁS Kft. nyomdaüzemében
Budapest, 1993. február hó