I.O.S.

THE EDITING OF UNDERWAY DATA ACQUIRED ABOARD THE R.R.S. "DISCOVERY"

by

WILLIAM K. STRUDWICK

I.O.S. REPORT NO. 22

1976

NATURAL ENVIRONMENT RESEARCH COUNCIL

INSTITUTE OF
OCEANOGRAPHIC
SCIENCES

THE EDITING OF UNDERWAY DATA ACQUIRED ABOARD THE R.R.S. "DISCOVERY"

by

William K. Strudwick

I.O.S. REPORT NO. 22

1976

I.O.S.,
Wormley,
GODALMING,
Surrey,
GU8 5UB.

Abstract

This paper attempts to synthesise the current problems and solutions associated with the underway data collected by the IBM 1800 computer aboard the R.R.S. "Discovery". It follows in detail the path of faulty data from its causes and creation through its detection and ultimate obliteration.

A semi-automatic system of data editing is described.

Introduction.

Those of us who, for our various reasons, have had course to use data collected by the IBM 1800 computer aboard the R.R.S. "Discovery" have often found to our dismay that data we had hoped to be complete and free from error was in fact infested with a host of bad or faulty readings. The reasons behind these errors are not immediately apparent: the instruments are not in themselves prone to inaccuracy and do not of their own volition feed faults into the computer.

It is within the computer and with equipment extraneous to both computer and the recording instruments that the problem lies. The computer software must by its complexity fail occasionally and it is these failures which cause some of the bad data and all of the temporal disorder.

Yet there is another problem which has hampered the production of clean data: the human factor. Even now, six years after the installation of the computer, there are still only a handful of people who understand the data in sufficient detail and who have the time and patience to qualitatively analyse it, to carefully compare various instruments reading and to give a well informed and accurate assessment on it. Only data which has every reading checked can be of any use, but to check every reading is a time consuming and often unrewarded pursuit.

Such is the current state of the art of data editing that the reliability of data is sometimes equated with the reliability, prestige or character of the person who edited it. That this state continues to exist is intolerable and yet little, or nothing, has been accomplished to alleviate it, until now.

With these ideas in mind I therefore undertook the task of devising a way of more accurately analysing data to determine good from bad and, ultimately, to produce an automatic system of data editing which would not rely upon the interest of an over worked or under-trained observer.

The Origins of Errors.

There are three types of error in "Discovery" data. These are caused by:-

(1) Interference of a recording instrument by the ship's radio equipment.

(2) Erroneous times and days produced by the computer.

(3) The packing of CDAT.

Each of these causes produce a distinguishably different type of error as will be discussed later.

At the present time of writing it seems unlikely that anything can be done to alleviate the interference caused by the ship's radio. Filters were introduced sometime ago to reduce the effect but they proved useless.

The phenomena of radio interference has been well studied (most recently by Andrew Bicknell and the author in 1974). The pattern of interference is unique for each broad spectrum of frequency. The result of a 12 MHz broadcast is quite distinct from those of 8 MHz and 6MHz. These differences can be seen clearly in the plots of one second data from the solarimeter (figures 1, 2 and 3), actual frequencies used are shown in Table 1.

In some instances it might be possible to ascertain a good value for a particular variable during a period of radio noise.

Table 2 is a print-out of one second solarimeter data during the night. Since the solarimeter values are usually around zero after dark the radio noise is clearly distinguishable as short bursts of high values. If it were of vital importance to know the real value of the solarimeter reading it could be obtained by rejecting values greater than ten and averaging the remaining numbers.

Unfortunately not all instruments are as convenient as this and to collect, store and analyse one second data for several instruments over a period of maybe twenty minutes would be a demanding and complex task. Apart from this, the results would be open to a good deal of criticism as to their accuracy.

In general, therefore data collected during radio transmissions must be rejected unless it can be proven that the interference caused an undetectable change in instrument readings.

The two other types of errors are unequivocally the product of software. Work on these problems is still progressing but positive results remain elusive.

It was discovered in 1974 that all timing errors were a product of one of three conditions which were:-

i) The end of a satellite pass.

ii) A system reload.

iii) A system cold-start.

However, the problem does not always occur when one of these events happens.

Determination of Errors.

It is usually easy to detect errors by reference to a plot of the variable against time. The errors defined as (1) and (3) can be clearly seen on figure 4 (reference 1). The "bumps" just after 8 and 12 hrs are radio noise; the dip at 0840 is the packing of CDAT and the disturbance at 1340 is cloud. This graph shows clearly four unusual areas in an otherwise smooth curve. One of these areas contains genuine data (being caused by slight cloud) but the others are definite errors.

However, in some instances it is virtually impossible to determine whether data is good or bad merely by looking at a single trace. Reference must be made to other instruments to see if an error occurs on these as well. Quite often an error can be followed through on all the instruments as is shown in plots of the four thermometers in figure 5. Notice that in the plots of the dry-wet thermometers the kink just before nine o'clock does not appear. This is not indicative of no error, rather that the error is very small for these instruments.

Errors during the night can be seen easily from a plot of the solarimeter since this instrument should read zero during the hours of darkness. During the day, however, the solarimeter readings are highly erratic and usually do not conform to the almost perfect trace shown in figure 4.

As was mentioned above each type of error has its own distinguishable features. Radio interference appears as a series of readings radically different from those which proceed and follow them. The effect of CDAT being packed is one or two readings which are different from the trend. Timing errors can appear as either of these.

The meteorological instruments have proved invaluable in discovering areas of faulty data. It has also been found that once time limits have been set for one instrument they often apply to all the instruments.

Accurate times for periods of faulty data cannot, of course, be obtained from plots. Exact times must be gained from listings of data otherwise there is always a chance of losing readings. Changes in tabulated data are easily fround once a general area of interest has been determined.

The classical approach to data editing.

Although the procedure for data editing varies slightly from person to person the approach laid out here is essentially that employed by Andrew Bicknell on "Discovery" cruise 68 (reference 2). Data was edited at sea on a daily basis.

There are several advantages to editing data within a day of its capture. For example, any anomalous values can be checked by reference to other sources of information, such as the ship's charts and log for navigation and people's memories for weather conditions.

The procedure was as follows:-

(1) The CDAT file is checked for erroneous records. If any are found then the day or times are altered so that time is always increasing or remains the same. (This is done so that another program may be run later to remove records of the same time).

(2) Produce plots of the following against time.

(a) Solarimeter.

(b) Corrected wind direction.

(c) Corrected wind speed.

(d) Port dry-wet bulb.

(e) Port dry bulb.

(f) Starboard dry-wet bulb.

(g) Starboard dry bulb.

(h) Course made good.

(i) Speed made good.

(j) Magnetometer (when available).

(3) Scrutinise these plots for data spikes and determine approximate times.

(4) Obtain a print-out of data between defined time limits.

(5) Determine exact times of erroneous data.

(6) Edit data by changing its status. In the case of corrected wind, north and east component velocities must also have their status changed. In the case of course made good and speed made good the em-log velocities must have their status changed.

(7) If em-log values have been altered then the corrected course must be re-computed to allow for the changes.

Seven programs were required for the above procedure, they are:-

(1) CHECK - this checks for the following types of errors:-

(a) Times the same.

(b) Times go back.

(c) Invalid day number.

(2) SYKEY - edit words within a record of a FORTRAN disk data file.

(3) METPL - plot starboard and port dry and dry-wet air temperatures, corrected wind speed and direction and solarimeter against time.

(4) GEOP1 - plot course and speed made good, distance run and depth against time.

(5) CDLPI - lists selected data paths from CDAT.

(6) STAED - edit status of given data paths in CDAT.

(7) ONAV1 - correct navigation.

The approximate times for each of these programs and the total time for the whole procedure is shown in Table 3.

Navigation Editing.

So far we have only considered data as a continuous variable related to time. But for the data to be of any use it must be associated with a geographical position.

The position of the ship is determined by the use of three instruments:-

i) An accurate positioning device (usually the Satellite receiver).

ii) The electromagnetic logs.

iii) The analogue or digital gyro.

For good navigation one must have all three devices working together. It is useless if only two of the instruments are working. Consider figure 6. Assume positions 1 and 4 have been accurately determined by use of the Satellite navigator. From position 2 to position 3 data from electromagnetic logs and digital gyro was not available. Consequently between position 2 and 3 there are an infinite number of courses the ship might have taken.

There are three ways of dealing with this situation:-

i) The course can be dead reckoned forward from position 1 to position 2, backward from position 4 to position 3 and a gap left from position 2 and 3.

ii) The same as (i) except positions between 2 and 3 are linearly interpolated. (This is quite good so long as the gap is fairly short or there is a positive indication that the ship did not radically alter course).

iii) As assumed course is made for the data gap and the navigation update proceeds as if no data is missing.

Although at first number (iii) might appear the course of action most likely to aid erroneous positioning it probably provides the most accurate, and certainly the most pleasing, navigation, so long as not too many guesses are incorporated into the data.

With more than one gap between fixes it becomes increasingly more difficult to interpolate or guess data and the results become even more questionable.

Consider figure 6 again. Assume now that positions 1 and 6 are accurately determined by satellite and all other positions have been calculated from em-log and gyro. The data from position 3 to 4 is completely lost, there is no way of accurately placing it.

In this case we could do the same as (iii) above, and if the gaps were no more than 4 minutes each then this would probably suffice. For longer gaps it would be dangerous to assume anything unless there was direct evidence from another source of the ship's movements.

Automatic Data Editing.

1. The procedure described above for editing data, though effective, is not only very time consuming but also relies totally on manual intervention. It would be far better if an automatic or at least a semi-automatic procedure could be devised. Let us consider the three types of error again and decide a way of pin-pointing exactly their occurrences.

i) Radio Interference.

In 1975 1800 voltage reference sampling was inaugurated. This has proved to be a very accurate gauge for determining the periods of radio transmissions.

A typical day's plot shown in figure 7. The voltage reference digit value varies only slightly during the day but peak's perceptably during radio transmissions.

Using this data as a guide it was possible during cruise 74 to write a very simple program to test the value and if greater than a chosen threshold to set status of all radio-interfered data to zero. This program was then used to great effect in cutting out faulty data.

i) CDAT Packing Spike.

The spike caused by the packing of CDAT is more of a problem. There is no variable which peaks perceptable for this event.

It is possible, however, to set a value for many of the instruments which the difference of any two readings will not exceed in normal circumstances.

To find these values of maximum change we must analyse the data.

Certain variables do not change very much and these are the best to look at. Four variables were chosen for analysis, these were:

i) Port dry air temperature.
ii) Starboard dry air temperature.
iii) Hull temperature.
iv) Barometric pressure.

We will consider for the moment the port and starboard dry air temperatures. First the periods of faulty data were determined. The differences in readings which start and finish the periods are shown in table 4. From this table we can see that all the differences are in excess of modules 0.4 and the large proportion are greater than modules 0.7. Also, not surprisingly, a period of bad data started with an increase ends with a decrease and vice-versa.

Let us now look at the data devoid of the sudden peaks, and troughs. Table 5 is a frequency distribution of the differences.

From this we can see that a difference of modules 0.5 also appears here. This is unhelpful since it would be far more convenient if there was a large gap between the maximum changes for good data and for bad. It is for precisely this reason that analysing differences is of limited use for finding all the areas of faulty data.

The difference distribution does tell us however that the vast quantity of differences lie in the range ±0.1. The actual percentages for increasing ranges of difference are shown in table 6.

We know that the CDAT packing operation causes only one faulty reading of data and therefore is easily distinguished, as in the port and starboard dry air tempertures in figure 5, by a sharp rise and fall (or fall and rise).

We must now make an assumption to aid the detection. Let us assume that the maximum difference for the port and starboard dry air temperature is modulus 0.4. If the difference is greater than this then we can place the further constraint that the difference which follows this one must be highly suspect if it lies in the range $-(d \pm 0.1)$, where d is the previous difference.

Using this mthod it is a simple task to write a program to look for these sudden peaks and to automatically edit the status of the data.

This method will not always work since the value entered into CDAT by the packing operation is a reading taken the previous day. If the readings are fluctuating over a long period of time then a noticeable spike will occur, if there is only slight variation the spike may not be so easily detected.

iii) Erroneous Times.

The automatic detection and correction of erroneous times is a very dangerous procedure to employ numerous types of errors that can actually occur render the task very complicated and from the work carried out on the problem it would still seem advisable to check times manually.

Automatic Programs.

2. The two automatic programs described above for data editing should not change the status of all data since some variables are less prone to radio noise than others and a few are totally independant of noise. The variables which do not have to be edited are listed below:-

i) Magnetic field (this is sampled once every seven seconds so is less prone to noise).

ii) Magnetic anomaly (as above since this is described from the field).

iii) 1800 reference voltage (this remains as a guide to the data that has been edited).

iv) Uncorrected depth.

v) Corrected depth.

vi) Matthew's area.

vii) Distance run.

The New Procedure.

3. We can now define a new procedure for the editing of data.

i) Check times.

ii) Run program to edit radio noise.

(iii) Run program to edit CDAT packing spike.

(iv) Produce plot of selected data.

(v) If CDAT packing not edited then edit manually.

(vi) Rerun navigation for edited periods.

Note that (iv) is required only because of the possible failure of (iii). Any variable could be plotted but the following have proved useful.

(i) Barometric pressure.

(ii) Starboard dry air temperature.

(iii) Port dry air temperature.

(iv) Hull temperature.

It can be seen from the above that the editing of data is now a very simple task and can be carried out by people of even limited experience. The whole system can be reduced to a very simple level by judicious programming: for example the edit programs can produce appropriate data for direct input to the navigation suite.

Table 3 can now be ammended and this is shown in table 7. Two programs need explanation:

(i) SPEAK -edits the radio noise.

(ii) PLOTS - plots the selected data.

The implementation of this new procedure should ideally contain some means of keeping track of what has been edited.

Conclusion.

The new procedure described above was developed on cruise 74 (September/October 1975) and worked very well. There is, however, still some work still necessary to ensure that the procedure runs smoothly and is effective. It is imperative that alterations made to the data, whether automatic or manual, are recorded by the computer and a detailed list kept for future reference.

The importance of manual quality control cannot be over emphasised; there are many occasions when an automatic suite becomes unreliable. For example, the computer might reject a sudden gust of wind of 100 knots when other readings are only 20 knots, but the one reading of 100 knots, if it is true, is of far greater significance than the other data.

It is essential that everyone who goes to sea with the IBM 1800 computer is aware of the importance of good data and the need to edit it.

Acknowledgements.

The author is grateful to numerous people for the assistance they have given both during and before the writing of this paper. In particular I would like to thank the following: Roger Hammerton, Radio Officer on Discovery, for the tremendous effort he put into keeping accurate records which helped so much in the analysis of radio noise; and Andrew Bicknell, who on cruise 68 proved beyond doubt the efficacy of editing data on a daily basis.

Also I would like to thank the scientists at IOS (Wormley) who have used "Discovery" data and have expressed interest in its quality.

References.

1. Bicknell, A. 1974. Unpublished data plots from cruise 68.

2. Bicknell, A. and Strudwick, W.K. 1974. The real-time editing of underway data collected on-board the R.R.S. "Discovery".

TABLE 1.

Frequencies used during the time periods represented by figure 1, 2 and 3.

DAY	TIME START (GMT)	TIME FINISH (GMT)	FREQUENCY (kH)
325	2007	2015	6275
326	2015	2025	8367
	2025	2035	8407
323	2010	2015	8367
	2015	2023	12550
	2023	2033	8367

TABLE 2.

List of one second solarimeter data for day 322 2017 hrs. on "Discovery" cruise 68 1974.

-6	-2	406	394	-6	414	-2	414	382	-2
-2	-4	438	406	398	-2	0	446	430	-2
0	-4	350	382	0	-4	-2	414	414	-2
414	238	-2	-2	0	2	0	-6	382	414
414	302	0	-2	-2	366	334	222	-4	-4
334	0	0	2	-2	374	-2	406	458	414

TABLE 3.

Timings for data editing steps.

Operation	Run time (minutes).	
CHECK	1	Computer time
SYKEY	5	
METPL	20	
GEOP1	10	
CDLPI	10	
STAED	10	
ONAV1	15	
Preparation of Papertapes and think time.	45	Man time
TOTAL	116	

TABLE 4.

Start and finish differences for bad data.

	Port	Starboard
Start	0.5	0.8
Finish	-1.0	-1.2
Start	-0.8	-0.9
Finish	0.8	0.9
Start	0.7	1.2
Finish	-0.8	1.4
Start	2.7	2.4
Finish	-1.6	-1.6
Start	-1.5	-1.3
Finish	1.3	-1.0

TABLE 5.

Difference distribution for good data only

Difference	Port	Starboard
-0.5	0	1
-0.4	0	0
-0.3	1	4
-0.2	5	13
-0.1	34	46
0.0	160	119
0.1	48	43
0.2	2	22
0.3	0	1
0.4	0	1

TABLE 6.

Percentages for increasing ranges of difference.

Range	Port	Starboard
0.0	64.0	47.6
± 0.1	96.8	83.2
± 0.2	99.6	97.2
± 0.3	100.0	99.2
± 0.4	-	99.6
± 0.5	-	100.0

TABLE 7.

Timings for data editing steps using the new procedure.

Operation	Run time (minutes)	
CHECK	1	Computer time
SYKEY	5	
SPEAK	1	
PLOTS	1Ø	
CDLPI	1	
STAED	1	
ONAV1	15	
Preparation of Papertapes and think time	5	Man time
TOTAL	39	

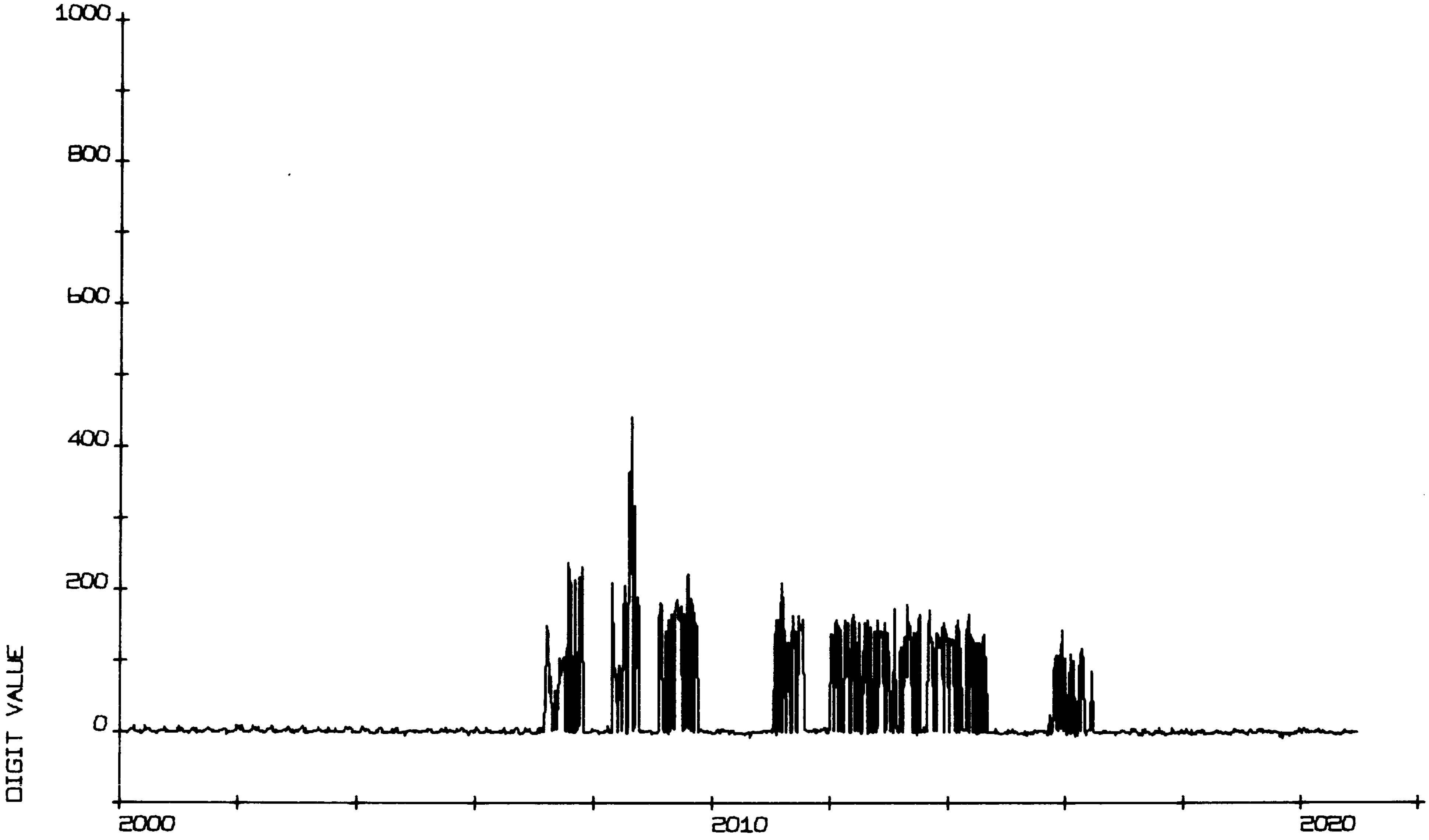

FIGURE 1. PLOT OF ONE-SECOND SOLARIMETER DATA FROM CRUISE 68 LEG 2
DAY 325 FROM 2000 HRS TO 2020 HRS

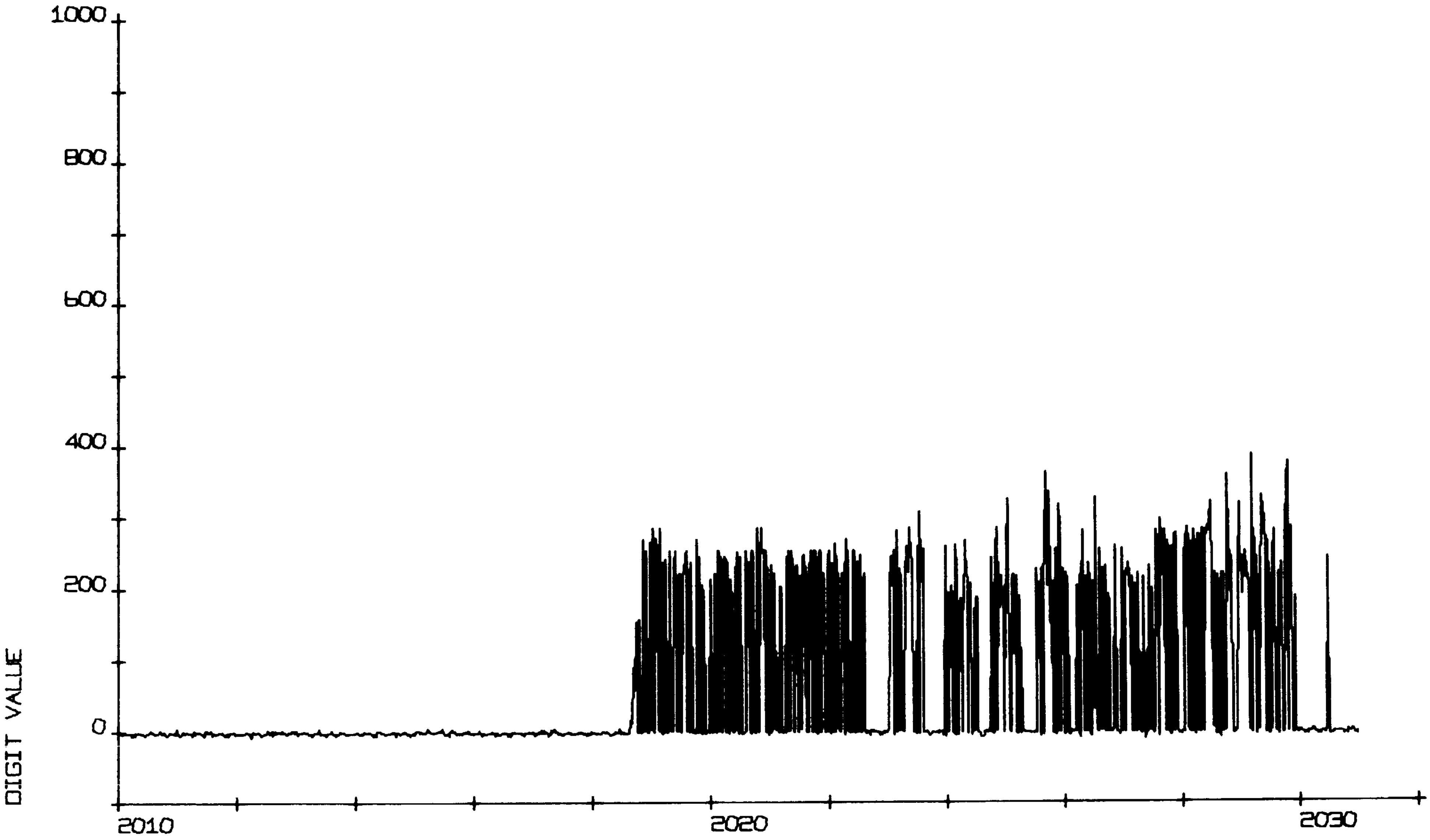

FIGURE 2. PLOT OF ONE-SECOND SOLARIMETER DATA FROM CRUISE 68 LEG 2
DAY 326 FROM 2010 HRS TO 2030 HRS

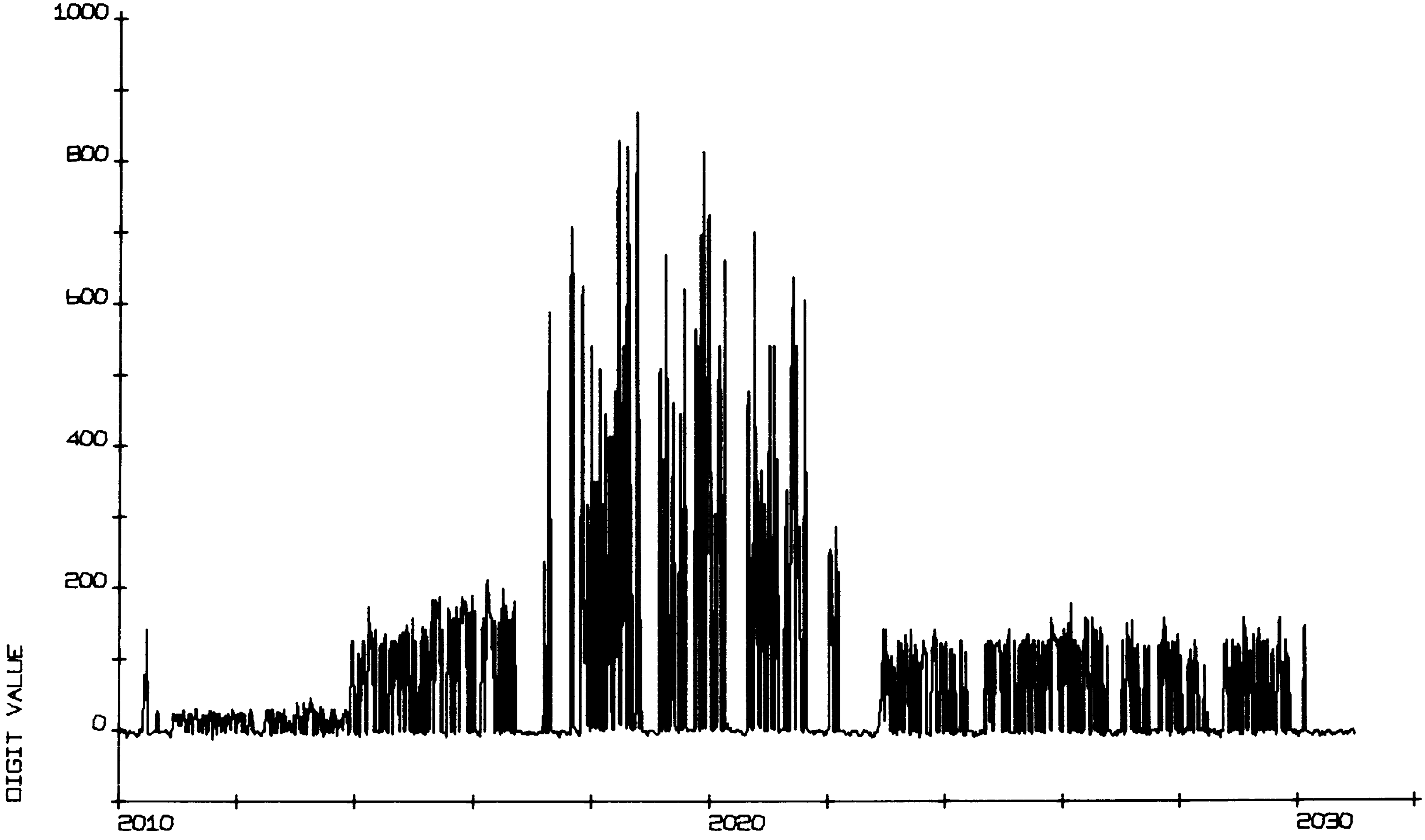

FIGURE 3. PLOT OF ONE-SECOND SOLARIMETER DATA FROM CRUISE 68 LEG 2
DAY 323 FROM 2010 HRS TO 2030 HRS

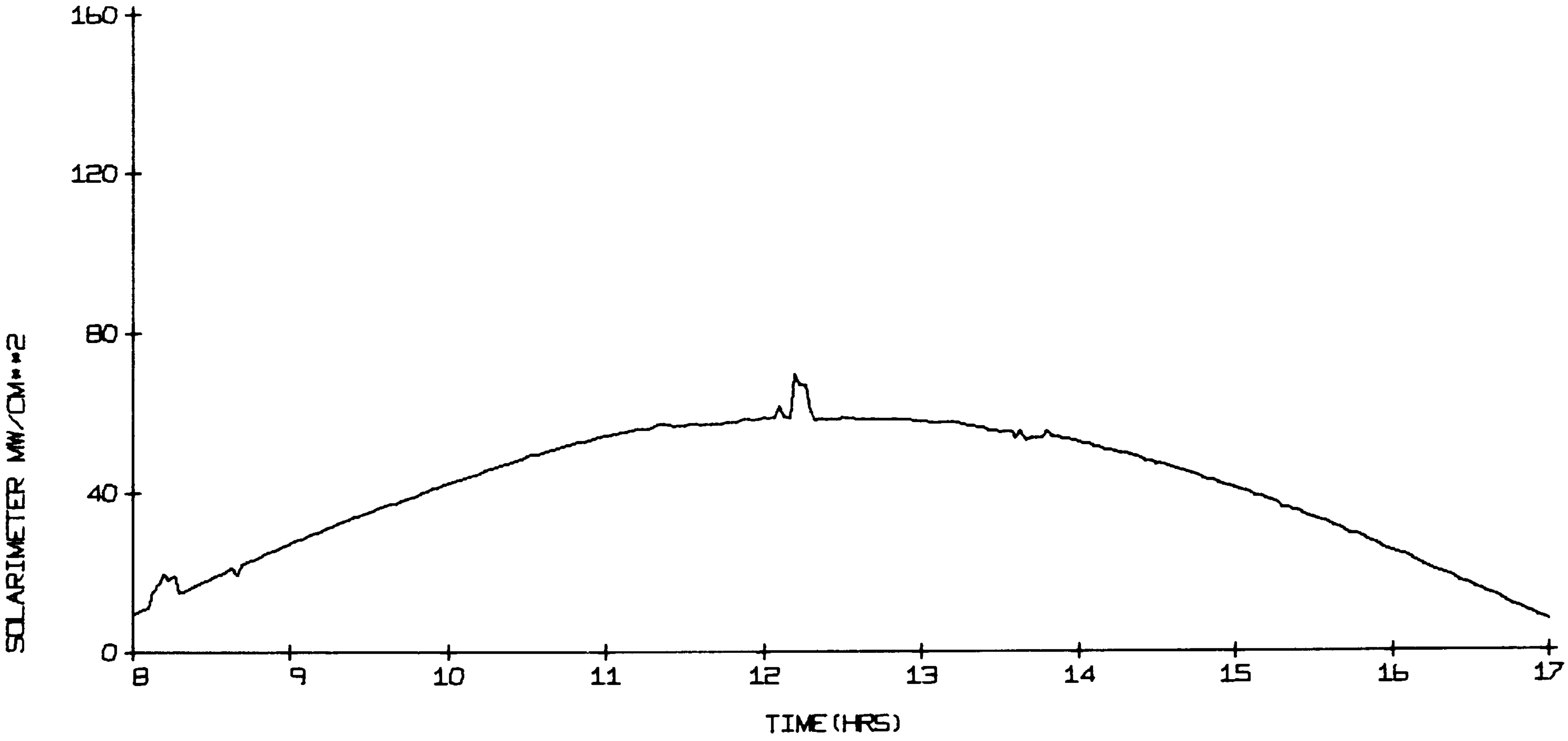

FIGURE 4. PLOT OF SOLARIMETER DATA FOR CRUISE 68 LEG 1 DAY 310

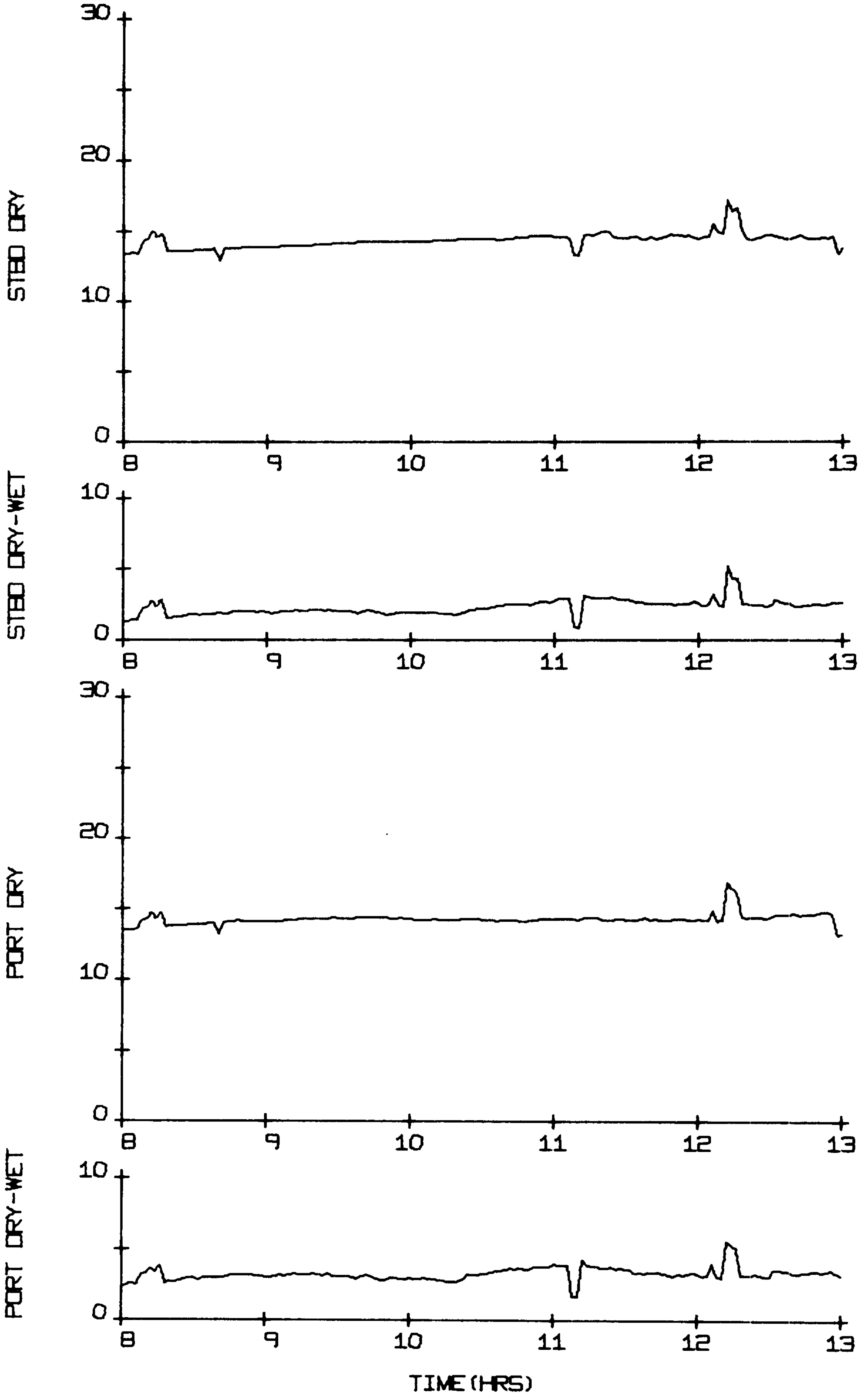

FIGURE 5. THERMOMETERS CRUISE 68 LEG 1 DAY 310

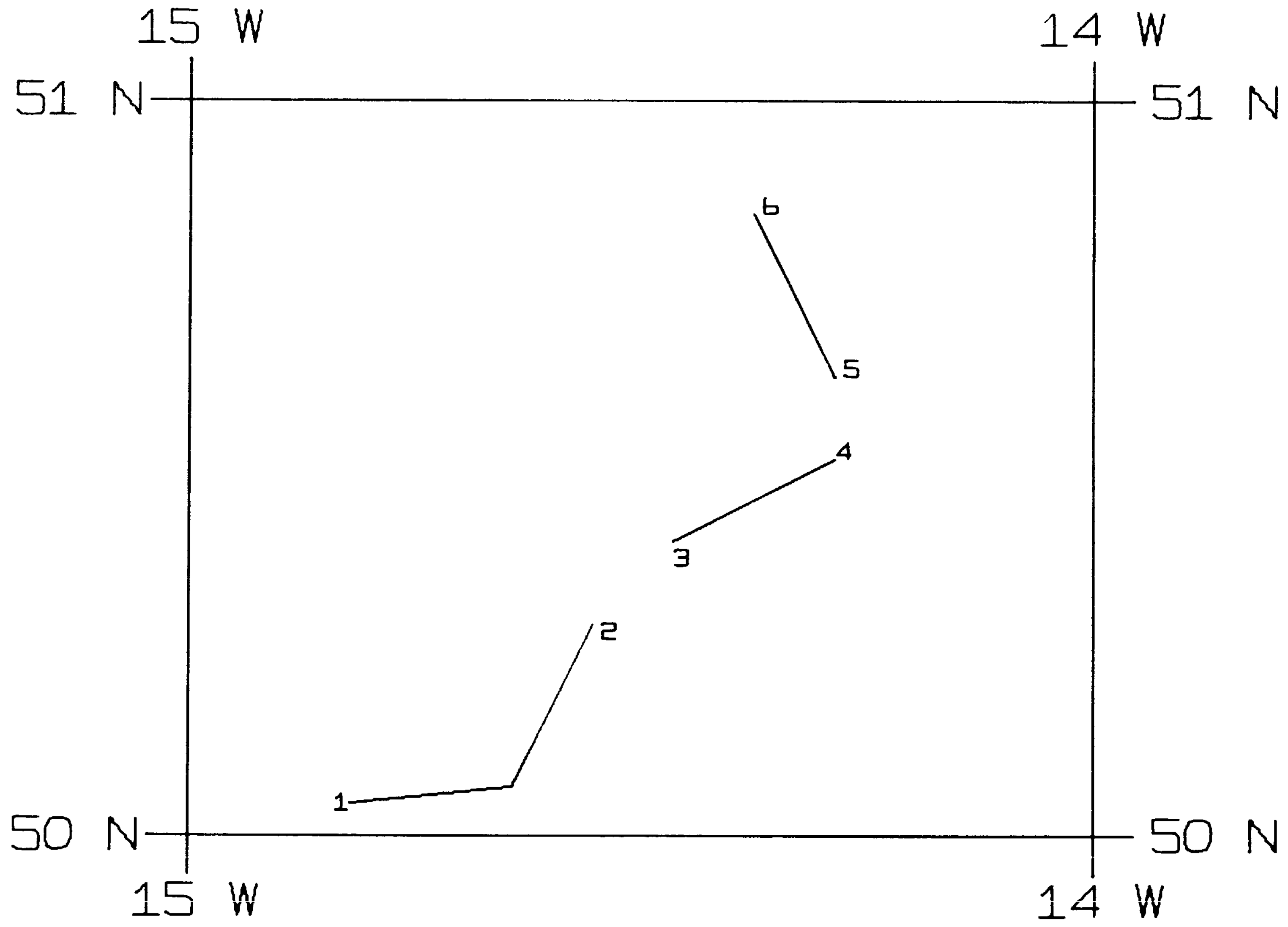

FIGURE 6. PLOT OF SHIP'S TRACK

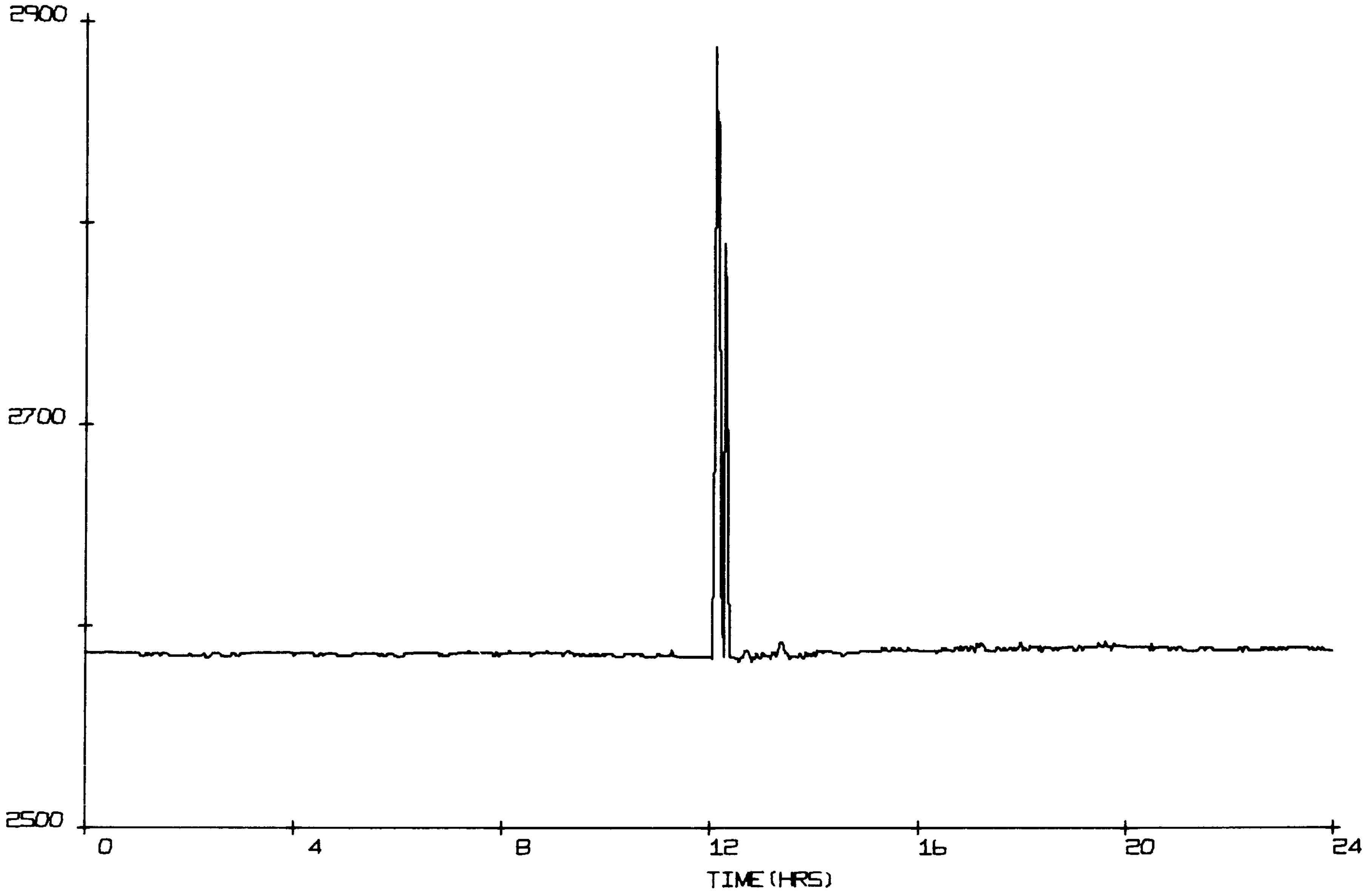

FIGURE 7. PLOT OF 1800 REFERENCE VOLTAGE CRUISE 74 LEG 3 DAY 276