AUTOMATIC CLASSIFICATION OF SEISMIC DETECTIONS FROM LARGE-APERTURE SEISMIC ARRAYS

by

SEYMOUR SHLIEN
B.Sc. McGill University (1968)

SUBMITTED IN
PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE
DEGREE OF DOCTOR OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June 1971 (i.e. JUNE 1972)

Signature of Author . . . .
Department of Earth and Planetary Sciences

Certified by . . . .
Thesis Supervisor

Accepted by . . . .
Chairman, Departmental Committee on Graduate Students

Lindgren
WITHDRAWN
MASS. INST. TECH.
JUN 20 1972
MIT LIBRARIES

Abstract

AUTOMATIC CLASSIFICATION OF SEISMIC DETECTIONS FROM LARGE APERTURE SEISMIC ARRAYS

by

Seymour Shlien

Submitted to the Department of Earth and
Planetary Sciences on May 5, 1972
in partial fulfillment of the requirements
for the degree of Doctor of Science

The large-aperture seismic arrays in Montana (LASA) and Norway (NORSAR) make on-line signal processing a necessity if these arrays are to be used at their full capability. Using the outputs of the detection processors of the respective arrays, the feasibility of automatic classification of seismic signals into the various body phases P, PKP, PcP, ScP, SKP, PP, PKKP and P'P' was confirmed. It was shown how these later phases can be used to advantage in improving the location capability using the combination of the two arrays.

One of the byproducts of this study was an estimation of the detection and location capabilities of the arrays. It was estimated that LASA detects more than 50 real seismic signals a day, of which less than 10% are due to later phases. LASA's detection capability extends almost one body wave magnitude below ERL's capability based on reported epicenters. The discrimination between very weak seismic signals and false alarms due to spurious noise was found difficult on the basis of only the detection logs.

Only a little more than 8 earthquakes a day were found common between LASA and NORSAR arrays. It is expected that this number will increase with the improved signal processing that the two arrays recently implemented.

Thesis Supervisor: M.N. Toksöz

Title: Professor of Geophysics

ACKNOWLEDGMENTS

I am indebted to Nafi Toksöz, who was my advisor throughout my four years of graduate study at M.I.T. He had originally suggested this problem and had given me constant encouragement.

I am especially grateful to Dr. Richard Lacoss, who had sacrificed a lot of his time in following my work. His sympathy for my frustrations and excellent sense of humor are deeply appreciated.

Thanks go to Dr. Charles Felix (IBM), Dr. Richard Lacoss (Lincoln Laboratory), Prof. Seymour Papert (Artificial Intelligence Group) and Prof. Michael Godfrey (Civil Engineering) whose discussions helped me get started. Also I would like to thank Mr. Larry Lande and Mr. Russell Needham, who had taught me to read seismograms.

In addition, I would like to thank Dr. William Dean of IBM, Mr. Simon Sarmiento (IBM), Mr. Albert Taylor (MIT), Mr. Lawrence Sargent (Lincoln Laboratory) and Miss Mary O'Brien (Lincoln Laboratory) who had helped procure the detection logs and summary bulletins.

Thanks go to Mr. Jerry Moore (IBM), Mr. Tom Murray, Mr. Philip Fleck and Miss Leslie Turek of Lincoln Laboratories, Mr. Richard Steinberg and Mrs. Jean Bow (Information Processing Center) for programming assistance.

Prof. Theodore Young (Pattern Recognition Group) had shown much interest and made many valuable suggestions in the final stages of this work.

I would also like to acknowledge the helpful discussions I had with Dr. Jack Capon and Mr. Robert Sheppard of Lincoln Laboratory and Mr. Guy Kuster and Mr. Norman Brenner (MIT).

I was supported by a Lincoln Laboratory Fellowship 1970-1971 and by a Chevron Oil Research Fellowship 1971-1972. This project was supported by the Advanced Research Project Agency and monitored by the Air Force Office of Scientific Research under Contract No. F44620-71-C-0049.

TABLE OF CONTENTS

Page

Abstract i

Acknowledgements ii

Table of Contents iv

List of Figures vi

List of Tables x

1. Introduction 1

2. LASA and NORSAR Capablities 8

    2.1 Introduction 8

    2.2 SAAC signal processor 8

    2.3 Detection capability of LASA and NORSAR 12

    2.4 Location capability of LASA and NORSAR 18

    2.5 Depth and magnitude estimation 19

    2.6 Conclusions 20

3. Pattern Recognition as Applied to Seismic Array Problems 23

    3.1 Introduction 23

    3.2 Pattern recognition 25

    3.3 Training set 31

    3.4 Summary 37

4. Classification of Detections Using One Array 38

    4.1 Introduction 38

    4.2 Single array false alarm discrimination 38

Page

4.3 The single array phase identifier 45

4.4 Conclusions 56

5. The Two Array Phase Identifier 59

5.1 Introduction 59

5.2 Two-array phase identifier 60

5.3 Locating earthquakes with two arrays 68

5.4 Conclusions 72

6. Conclusions 73

References 77

Appendices 80

A Criterion for Matching Predicted Signals to the Detection Log 80

B Distance and Azimuth Resolution of a Large-Aperture Seismic Array 83

B.1 Introduction 83

B.2 Distance resolution 83

B.3 Azimuth resolution 85

C Improved Discrimination of Signals from False Alarms 88

D Shuffling a Detection Log 91

E Numerical Evaluation of l_i for the Single Array Phase Identifier 93

F Spherical Surface Transformation 96

G Estimation of Λ_{ij} 98

LIST OF FIGURES

1. Ray paths of the different seismic phases (Richter, 1958).

2. LASA seismograms of large magnitude ($m_b \sim 5.5$) earthquakes at various distances (Δ) from LASA. The gain factor was adjusted for the best reproduction. The phases from the distant events have a low signal-to-noise ratio. The seismograms have been sampled 20 times a second. The bottom trace indicates two-second marks.

3. (Upper) Block diagram of LASA signal processor. (Lower) Block diagram of LASA's Detection Processor.

4. P & PKP phase locations of LASA beams (high resolution partition) as given by SAAC and plotted on an equidistant projection centered at LASA.

5. Number of earthquakes reported by ERL (white) & number of earthquakes reported by ERL and detected by LASA (black) as a function of distance.

6. Number of earthquakes reported by ERL (white) & number of earthquaes reported by ERL and detected by NORSAR (black) as a function of distance.

7. Percentage of ERL earthquakes less than 80 degrees from the respective arrays that were detected as a function of body wave magnitude.

8. Percentage of ERL earthquakes greater than 80 degrees distance from the respective arrrays that were detected as a function of body wave magnitude.

9. Number of earthquakes reported by ERL (white) and number of earthquakes reported by LASA (black) as a function of distance.

10. Correlation of LASA body wave magnitude versus ERL body wave magnitude for the same events.

11. Frequency magnitude distribution of earthquakes less than 95 degrees from their respective arrays determined from LASA and NORSAR bulletins and ERL catalog.

12. Time interval between later and first arrival in seconds measured at LASA versus ERL determined epicenter distance from LASA for PcP, ScP, PP-P PP-PKP and SKP phases.

13. Time interval between later and first arrival measured at LASA in seconds versus ERL determined distance of epicenter from LASA for P'P', PKKP-P, and PKKP-PKP phases.

14. Histograms of distance and azimuth errors of epicenter determinations for LASA and NORSAR Summary Bulletins.

15. Maximum MSTA in a detection group versus summary bulletin reported amplitude in millimicrons for LASA and NORSAR. Scatter was reduced by averaging MSTA over 0.1 millimicron units.

16. MSTA distribution of all detections and of only false alarms in the LASA and NORSAR detection log.
17. False alarm probability given MSTA for LASA and NORSAR.
18. Cumulative probability distribution function of the maximum MSTA of detection groups matched to reported signals in the LASA and NORSAR Summary Bulletins.
19. Cumulative number of reported events in summary bulletin, estimated cumulative number of false alarms, and estimated cumulative number of signals, versus MSTA for LASA and NORSAR.
20. Travel time interval between first arrival and later phase versus the inverse phase velocity of the later phase.
21. Empirical frequency distribution of parameters S_1, S_2, and S_3 of the LASA single array phase identifier determined for correctly identified detection pairs and random detection pairs.
22. Log MSTA ratio of fist arrival to later phase for various phases.
23. LASA (L), NORSAR (N), and chosen ($\underline{L}$) beam locations plotted with travel time interval curve for various phase combinations. Actual epicenter is at X.
B-1. Distance and azimuth of epicenters triggering specific LASA high resolution beams.
B-2. Standard error in azimuth versus SNR parameter a_o.

E-1. Cumulative probability distribution function of LASA's MSTA for all beams (left) and for only aseismic beams (right).

F-1. Partial derivative of distance of a point from NORSAR with respect to distance from LASA as a function of LASA distance and azimuth. Grid spacing is in 10 degree intervals and the origin is at (0,0).

LIST OF TABLES

1. Confusion matrix for single array phase identifier.
2. Distribution of correctly and incorrectly classified training phase pairs determined from the two-array phase identifier.
E-1 Approximations to the decsion parameters of the single array phase identifier.

CHAPTER 1

Introduction

The Large Aperture Seismic Array in Montana (LASA) has made it possible to detect and locate earthquakes in real time over at least half the surface of the earth. Through the on-line processing of signals from 525 seismometers spread over an aperture of 200 kilometers, noise has been reduced to low enough levels to multiply the number of detectable earthquakes by at least a factor of two. The Seismic Array Analysis Center (SAAC) at Washington reports about 30 earthquakes on an average day. These earthquakes are located within several hundred kilometers within several hours after they occurred.

LASA is generating a very large data base by which one can eventually map the interior of the earth to finer detail. This thesis is mainly devoted to studying the contents of the detection log. The detection log is the direct output of the Detection Processor (DP) which attempts to flag every signal arriving at LASA. Many of the detections are not signals but false alarms due to the noise level suddenly increasing. The signals consist of mainly seven different body wave phases. If these detections could be automatically classified, the load of the analyst could be reduced considerably in the preparation of earthquake reports.

Most of the signals detected by LASA are the first arrivals namely P or PKP depending on the distance of the earthquake from the array. In about 10 percent of the cases a later phase such as PcP, ScP, SKP, PP, PKKP or P'P' is also detected. Later phases are caused by reflections of the seismic signal off the earth's core or free surface. (See Figure 1.) These later phases are both a nuisance and a boon. If a later phase is mistaken as a P phase then a fictitious earthquake would be reported. On the other hand later phases permit one to get a better estimate of the earthquake's epicenter and may be a deciding factor in determining whether a detection is real or not. A statistical pattern recognition technique will be developed to classify these detections either using a single array, LASA, or using LASA in conjunction with the Norwegian Seismic Array (NORSAR) which went into full operation in March 1971.

The nature of seismic signals are very variable due to effects of source mechanism of the earthquakes and the various inhomogeneities along the ray path. Sample seismograms are shown in Figure 2. Because of this, it is not feasible to incorporate a standard waveform, and the pattern recognition scheme will probably not perform as well as an analyst who has all available information. Nevertheless, the automatic classification scheme will save the analyst a considerable amount of time and standardize the identifications. Eventually an analyst may be necessary to only verify the output of the automatic phase identifier and resolve any conflicting phase identification.

One of the byproducts of this study will be an estimate of the capabilities of NORSAR and LASA. Estimates of the detection and location capabilities are needed for the automatic phase identifier. Since the estimates obtained here are based upon pre-processed data, they will highly reflect the quality of the initial signal processing and will not be the maximum capabilities of the arrays. This became very evident after this analysis was performed when LASA and NORSAR upgraded their signal processing.

In this study, we had a very small standard data base. Very few of the detections could be identified by an outside source. It was necessary to rely very heavily on the earthquake catalog distributed by the Environmental

Research Laboratory (ERL) to identify some of the detections. Since the ERL catalog only reports a fraction of the world earthquakes, there was no way of ascertaining that a specific detection is a false alarm due to spurious noise. Furthermore for many cases it was very difficult to positively identify a detection using the ERL catalog. There was always an uncertainty whether a predicted phase was properly matched to the detection. For instance it is conceivable that the signal was too small to be detected by LASA and what was observed was either spurious noise or some other signal from a different earthquake. Since the set of pre-identified detections (which we shall later call the training set) was used both to develop and evaluate the performance of the automatic phase identifier, some of the analysis was a little subjective. There was unfortunately little choice in this matter since only three months of data was available.

The effects of very deep earthquakes were completely ignored in this study. Because 90 percent of earthquakes are relatively shallow and depth effects are complicatedly related to the phase identification and epicenter determination, they were not incorporated into the phase identifier. Generally it is very difficult to distinguish depth phases such as pP from the seismic coda without seeing the actual waveforms. For earthquakes shallower than 100 kilometers, the travel time corrections were usually less than thirty seconds and could easily be neglected.

Except for the Seventh IBM Technical Report (1970) there was nothing published on the phase identification problem. No elaborate evaluation on the performance of their scheme has been reported yet.

The remaining part of the thesis is divided into five chapters. In Chapter 2, the SAAC signal processing is described and the capabilities of the arrays are determined. The first section describes how the detection log is generated at LASA from the raw signals coming into the 525 seismometers. The beam partitions used by the detection processor is discussed. Off-line processing to generate the summary bulletins is very briefly described. In the next section the detection capability of the arrays is estimated as a function of distance and magnitude on the basis of the summary bulletins and the detection log using the ERL epicenter determinations as an outside standard. Since LASA detects many more earthquakes than are listed in any earthquake catalog we had to resort to frequency-magnitude distributions to infer the lower magnitude limit of LASA's detection capabilities. The second half of this section describes the location capability of the arrays and the factors that determine this capability.

In Chapter 3 the theoretical framework necessary to understand how the automatic phase identifier works is described. A model of decision making is discussed and the concept of a training set is introduced. The statistical pattern recognition technique is described in the next section and examples are given to relate this method to the problem of distinguishing false alarms from signals and classifying phases. Bayes rule and the maximum likelihood test is briefly reviewed. "A priori, a posteriori probabilities", "observation space" and "performance" are defined. An alternative rule which uses the concept of distance is introduced. The distance rule is equivalent to the maximum likelihood test if the decision parameters have an error which is normally distributed.

In Chapter 4 the automatic phase identifier which uses a single array is described. Distributions of the decision parameters are determined and approximated. The programming of the automatic phase identifier is discussed and the performance of the phase identifier is determined from the LASA detection log.

Chapter 5 describes the two array phase identifiers. Much more information is available from the combination of LASA and NORSAR detection logs so that 50 different interpretations for a pair of signals can be distinguished. The distributions of the two array decision parameters are determined, the programming is described and the performance of the identifier is evaluated. A method of improving the epicenter determined from the two arrays when later arrivals are found is described.

In the final chapter results of this study are summarized and conclusions are drawn.

Throughout this thesis an attempt was made to put all the details and mathematics into the appendices. This was done to make the text more readable.

The data analyzed in this thesis was confined to the time period May, 1971, to August, 1971.

Chapter 2

LASA and NORSAR Capabilities

2.1 Introduction

In this chapter we discuss the present LASA signal processors, their capabilities and limitations. We start with the detection of seismic signals, and follow this by an estimation of detection capabilities of LASA and NORSAR as a function of distance and magnitude, in Section 2.3. The location errors are determined in Section 2.4, and in Section 2.5 we discuss the problems of location errors and magnitude estimation.

2.2 SAAC Signal Processor

The signal processing described here is basically that of LASA which was designed and developed by IBM and which went in full operation as of April 1969. The details of the present signal processor are described in the IBM final report (1972).

The processing of the seismic signals by LASA can be separated into three steps:(1) detection processing (2) event processing and (3) verification. A block diagram is shown in Figure 3. Since the input of the automatic classifier is the output of the Detection Processor the first step is described in a fair amount of detail while the other steps are dealt with briefly.

A teleseismic signal arrives at the array as a plane wave with a specific velocity and azimuth depending on the location of the earthquake and the phase type of the signal. If the output of the individual sensors of the LASA could be combined to screen out all the signals except that coming with the specific velocity and from the specific direction, the Signal-to-Noise-Ratio (SNR) may be enhanced considerably. The first step of the Detection Processor (Figure 3) is to generate in real time a set of 600 presteered beams with different velocities and azimuths. The beams are formed by delaying and summing the signals of the individual sensors. Let $S_i(t)$ be the amplitude of the signal at the ith sensor positioned at $\bar{x}$. Let $\bar{v}_m$ be the velocity vector corresponding to beam m. Then the delay times $t_{m,i}$ for the sensor i and beam m is given by:

$$t_{m,i} = \bar{v}_m \cdot \bar{x}_i / |\bar{v}_m|^2 \tag{2.1}$$

The beam $b_m(t)$ is formed from the individual sensors using

$$b_m(t) = \frac{1}{N} \sum_{i=1}^{N} S_i(t - t_{m,i}) \tag{2.2}$$

where N is the total number of sensors used by the beam generator. The resolution of the beam Δp in inverse velocity space is proportional to T/A, where T is the period of the signal and A is the aperture of the array. On account of the configuration of LASA sidelobes are very considerable. The biggest sidelobe is only 5 db below the main lobe (IBM Final Report, 1972).

The 600 beams can be separated into two overlapping partitions of 300 beams each. The first partition which has been in operation since April, 1969, is the set of high resolution beams. Because these beams are narrow, a very large number of beams are needed to cover all possible areas of velocity space from which one can expect the seismic signal. For economic reasons only 300 of these narrow beams are computed. These beams were pointed towards the seismic regions and areas of interest to monitor nuclear explosions. A plot of these beams on a world map is shown in Figure 4 for the P and PKP phases.

It is evident that the fine beam pattern leaves many gaps in the signal space, in particular for some of the later phases such as PP. If a seismic signal comes from an area where there is no beam coverage it would be missed by the detection processor if it is a weak signal. However, if the signal is very strong it will leak into a sidelobe of a beam which is pointed very far from the actual signal source. Since this was found to be undesirable, another beam partition consisting of low resolution beams was added to the Detection Processor in January, 1972. The second beam partition covers all the seismic signal space, but has much less resolution. Similarly, NORSAR has a fine beam partition of 331 beams and a broad beam partition of 160 beams.

Each of the beams is filtered and rectified by the Detection Processor. The filter was designed to deemphasize those frequency components where the SNR is low. In the case of the LASA array the signal is confined to a narrow band 1 Hz. The signal at NORSAR covers a broader frequency band. The rectified beams then pass through two integrators of different time durations. These integrators compute a Short Term Average (STA) and a Long Term Average (LTA). The LTA is determined over a 32 second interval and is supposed to be a measure of the natural noise. The STA is computed for a 0.8 second time interval and is a measure of the amount of signal if present. Both of these averages are updated every 0.8 seconds for all the beams. If 20 $\log_{10}$(STA/LTA) is above 8 db for at least two seconds, then the particular beam is declared to be in the detection state.

A large signal will usually trigger several beams simultaneously. The beams with the maximum STA in each of the beam partition are recorded onto the detection log for that particular time cycle. A large seismic signal usually has several bursts of energy so that as many as 15 beam detections could be recorded for just a P phase.

The LASA detection log contains 500 detections on an average day. Many of these are false alarms. The Event Processor (EP) searches through the log for signal detections with a large SNR and processes these signals off-line to refine the estimates of the signal amplitude, velocity and arrival time. The best fitting plane wave is found by a sequential,iterative, cross-correlative procedure (Farrell, 1971). Assuming the signal is a P wave, the epicenter, origin and magnitude of the earthquake can be determined from these parameters. The Event Processor reduced the number of possible signals to around 60 for an average day (Mack, 1971, personal communication).

The output of the Event Processor is next carefully screened by trained analysts. The analyst checks that the delay times of the subarray traces have been determined accurately and that the signal is indeed a P phase and not a depth phase, or a later phase or a "glitch." After making the corrections and recomputing the epicenter if necessary, he compiles the summary bulletin report which is distributed two days later.

2.3 Detection Capability of LASA and NORSAR

Detection capabilities of seismic instruments are bounded by the natural noise. There are many sources to microseismic noise; the natural sources are wind action, ocean waves and storms (Lacoss, et al., 1969). Man-made noises are generated by mining operations, trains, planes, etc. NORSAR has a much higher background noise level than LASA since it is situated much closer to the coast (IBM Final Technical Report, 1971).

With a large array of seismometers it should be possible in theory to reduce the noise to levels lower than observed by any single seismometer. Assuming the signal is coherent and the noise is independent from sensor to sensor, the SNR is multiplied by $\sqrt{N}$, where N is the number of sensors. Thus, with an array of 525 seismometers, the gain of SNR should be 25 db. Actually, this figure is an overestimate, since the noise among nearby sensors is not spatially incoherent while the signals between remote sensors are considerably different. LASA obtains a gain in the range of 10 to 15 db. Signals as small as 0.3 millimicrons are reported routinely by the LASA summary bulletin. This event would be only visible on a properly directed beam.

The detection capabilities reported here are not the ultimate capabilities of the arrays, but are representative of the Detection Processor's capabilities. Station correction used in beamforming are being upgraded as more data becomes available. Both LASA and NORSAR had some incorrect station corrections incorporated in their beam patterns when this analysis was made. Substantial improvements have occurred since some of these errors have been fixed.

Using the ERL preliminary epicenter determinations as a standard, the capabilities of the LASA and NORSAR arrays were estimated by counting the number of matches that could be made with the detection logs against the ERL catalog. The criterion of determining a match is discussed in Appendix A. The number of expected matches and observed matches as a function of the distance of an earthquake from the array are shown in Figures 5 and 6 for LASA and NORSAR, respectively. Periods when the Detection Processor was down were taken into account. In general, LASA detects more than 80% of the ERL events in the distance range between 20 and 90 degrees. These events were also listed in the Summary Bulletin. The NORSAR does not perform as well. The percentage of matches in the same distance range is down to 60%. The anomalous low number of detected events near the 60 degrees apparently are due to bad station corrections in the beams. These events are mainly in North America, Japan and Aleutian areas.

Local earthquakes (less than 20 degrees) cannot be easily detected or located by any array. The signal is usually emergent and spread out in time and the wavefront cannot be approximated by a plane wave. The $dT/d\Delta$ has such a large variation that a prohibitive number of beams would be needed to cover the signal space.

The P wave becomes diffracted by the core-mantle boundary beyond 90 degrees and its amplitude decreases rapidly with distance. Low magnitude earthquakes tend to be missed beyond these distances. In the diffraction zone the velocity of the P wave becomes independent of distance. Hence it is very difficult to locate earthquakes from this zone. LASA does not attempt to report any events beyond 100 degrees.

The detection capabilities were next estimated as a function of body wave magnitude. Events were separated into two groups, those less than 80 degrees from the array in question and those greater than 80 degrees. The fraction of detected events were determined as a function of magnitude for both groups and were plotted in Figures 7 and 8. Due to the small sample sizes at the higher magnitudes the ratio sometimes decreases with magnitude. The differences in the detection capabilities between the LASA and NORSAR arrays are now very apparent. Less than 20 percent of the ERL events in the distance range 20-80 degrees and in the m_b range 3.5-4.0 were detected by NORSAR. On the contrary, LASA is able to detect more than 80% of the events in this range. It is expected that NORSAR will improve its detection capabilities once better station corrections are incorporated into the Detection Processor.

Unlike LASA, a considerable amount of signal energy in the NORSAR is high frequency. The effect of poor station corrections is much worse. If two subarray traces are misaligned by a fifth of a second, a nontrivial fraction of the signal energy is lost. Station corrections for the NORSAR are just as large as for LASA. (Sheppard, 1971, personal communication). Station corrections reach values of 2 seconds for vertical incident waves at LASA. They are believed to be caused by the corrugated structure of the Moho (Greenfield and Sheppard, 1969). The spatial coherency of the signal is not any better for NORSAR. For these reasons NORSAR will never reach the same performance level as LASA.

The LASA Summary Bulletin reports many low magnitude events that are not in the ERL earthquake catalogue. In Chapter 4 it shall be inferred indirectly that LASA probably detects 50 earthquakes a day. About 30 events a day are listed in the LASA Summary Bulletin. In Figure 9 the number of earthquakes reported by LASA is plotted along with the number of earthquakes reported by ERL for the same time period, May to August, 1971, as a function of distance. The fact that the LASA seismicity distribution highly reflects the ERL seismicity distribution after taking into account the places where LASA is less sensitive to events, almost confirms the LASA reported events (LASA stops reporting earthquakes at around 95 degrees).

In order to get a better estimate of LASA's detection capability, the frequency-magnitude distribution of events reported by LASA was determined. (The correlation plot of LASA m_b estimate versus ERL m_b estimate in Figure 10 implies that the LASA body wave magnitude estimate is unbiased). Figure 8 plots both the LASA and ERL frequency-magnitude distribution for the same time period. ERL events further than 95 degrees from LASA were not included in the distributions since SAAC reports virtually no events beyond this distance.

The right hand portions of these distributions are in accordance with Richter's log-frequency-magnitude relation ($\log N = a - bm_b$), (Richter, 1958). From local micro-seismicity studies in which sensors are located within tens of kilometers of the epicenter regions, it may be safely assumed that Richter's relation extends to zero magnitudes. Both the tendency to ignore weak local earthquakes and background noise levels preventing the detection of weak teleseismic events causes the frequency-magnitude distribution to reach a turning point. The magnitude of this turning point is a very good indication of the detection capability of a network of array of seismometers. It was found by this method that ERL's detection capability is not uniform all over the world. For North and Central America the turning point was found to be around $m_b = 4.2$, while for western China, Indonesia, Australia area the turning point was at $m_b = 5.0$ (Shlien and Toksoz, 1970a).

It may be concluded from Figure 11 that LASA detects earthquakes down to a magnitude of 3.7. There appears to be a whole magnitude difference between the turning points of the ERL and LASA distributions. A small part of this difference may be due to a magnitude bias of LASA versus ERL which is very difficult to estimate by any conventional statistical method. ERL only lists about half of the earthquakes that it detects. If the reporting stations are too few in number or poorly distributed so that no location accurate to within a few degrees can be made, ERL will usually ignore this event (Sheppard, personal communication). Furthermore, there is a tendency to regard weak events as unimportant. A similar frequency-magnitude distribution based on the NORSAR Summary Bulletin is shown in Figure 11 for a comparison. Since March 1972 NORSAR has been reporting about twice as many events. Unfortunately, insufficient data was available at the present to justify repeating this analysis.

2.4 Location Capability of LASA and NORSAR

The location capability of an array depends on its resolution, accuracy of station corrections, and the distance of the earthquake from the array. A theoretical analysis of these factors is given in Appendix B.

In this section, the performance of the arrays in locating the epicenter of an earthquake is estimated on the basis of their summary bulletins and the ERL catalogue. It is assumed that the ERL epicenter is an accurate and unbiased estimate. Figures 12 and 13 justify this assumption. The travel time interval between the P phase and any later arrival is very sensitive to the distance of the earthquake. In these figures the time interval between these phases measured at LASA is plotted with respect to the distance using ERL epicenter determinations. No compensation for the depth of an earthquake was made. (For the time interval between phases, these corrections are usually less than 15 seconds for all but a few earthquakes). The degree to which the data points for the PcP and PKKP phases define the travel time curves attests to the accuracy of the ERL location. The other phases such as PP and ScP have longer periods and are emergent. Hence, their onset times could not be determined accurately. Points completely off the travel time curve are probably due to a phase misclassification.

In Figure 14, the distribution of the distance and azimuth errors are plotted for LASA and NORSAR. Large errors were generally caused by events near the shadow zone of the P wave (90 degrees and greater from LASA). Azimuth errors are generally very small for LASA. LASA has been locating epicenters for the past five years so it is probably operating at its ultimate capability. For NORSAR the errors in distance are generally larger. Some large distance errors were found for events 60 degrees from NORSAR (Aleutian and Japan areas) in addition to the shadow zone. The errors in azimuth are very bad. The large bias should disappear once NORSAR has been running for a longer time and the station corrections are improved. It is not believed that NORSAR is performing to its full capacity. The newer data is expected to have considerably smaller mislocation errors.

2.5 Depth and Magnitude Estimation

The depth of an earthquake can be determined with a single array only if the depth phases such as pP or sP can be found and distinguished. Because depth phases can be confused with the PcP phase or with just part of the P wave coda, this method is not reliable except for the rare clear-cut cases. The verification of a depth phase is done by comparing the actual waveform with the initial phase. The depth phase should be almost identical to the initial P phase except for a 180 degree phase shift. Automatic methods using spectral correlation methods were found to be unreliable by SAAC. SAAC no longer publishes the depth determination of earthquakes.

Due to the sensitivity of signal amplitude to many factors such as the structure underneath the array, $d^2T/d\Delta^2$, source mechanisms, and inhomogeneities along the ray path such as dipping plates, a single station or array cannot hope to estimate magnitude to more than an accuracy of half a unit. The correlation of LASA and ERL body wave magnitude estimates in Figure 10 showed the typical scatter found in any such investigation.

2.6 Conclusion

Large-aperture seismic arrays have extended our detection capabilities to new levels. Reliable earthquake bulletins covering most of the world can be put out within several hours. However, an array cannot compete with a large network of seismometers in locating earthquakes. With a network of stations suitably spaced around the earthquake, location is basically determined using travel time and spherical geometry. The arrival times of an earthquake phase at four stations contains sufficient information to estimate the latitude, longitude, depth and origin of the event. With a single array location can only be determined from the derivative of travel time with respect to distance. The estimation of this derivative is based upon the measure of the delay times of the seismic signals at the different sensors. The delay times generally do not exceed 25 seconds between the extremeties of the array. Individual station corrections run as high as two seconds and are very sensitive functions of the distance and azimuth of the earthquake epicenter.

The determination of these station corrections requires a set of accurately located earthquakes. An array must be calibrated before it can publish reliable bulletins. Various attempts have been made to develop models of the structure underneath LASA in order to explain these station corrections and amplitude variations. (Larner, 1970), (Greenfield and Sheppard, 1969). The amplitude of the seismic signal varies by almost an order of magnitude between sensors. Though these amplitude variations are repeatable for earthquakes coming from the same area, the pattern of this variation changes very dramatically and unpredictably as the epicenter moves several degrees. The modeling of the structure underneath LASA is complicated further by the highly irregular spacing of the seismometers. The seismometers are heavily concentrated near the center of the array and become very sparce towards the extremeties. The simple crustal structure used generally gives a gross approximation to the observations. The actual structure is probably very complex.

The signal variations across the array appear to be caused by multipathing. Mack (1969), showed that the seismic signal arriving at a single subarray is the result of many closely spaced individual arrivals which interfere with one another. He asserts that the multiples do not appear to be generated by a reflection process but rather by a wave-splitting phenomenon and diffraction. To be able to run LASA or NORSAR at their maximum capabilities these effects would at least have to be known if not understood, and much more complicated signal processing would be involved.

Chapter 3

Pattern Recognition as Applied to Seismic Array Problems

3.1 Introduction

The goal of the remaining part of this thesis is to develop an automatic classification scheme which will find the best identification for each detection in the detection log. Detections can fall into many different categories, the major ones being signal and false alarm. The signals can be subdivided into the different short period phases observed at LASA viz P, PKP, PcP, ScP, SKP, PP, PKKP and P'P'. For purposes of simplification, depth phases have been completely ignored. They would tend to be identified as their corresponding phase, thus pPcP would be identified as PcP. A few other phases such as SKKP and PKKKP are occasionally observed. They were also ignored on account of their rarity.

The input information for distinguishing signals from false alarms is much different than the information for classifying the signals. For this reason they shall be treated as two separate problems.

In almost all cases it is impossible to identify a signal as a particular phase without additional information. For the single array case the analyst identifies a later arrival by its context. Except for the shadow zone it is very unlikely for the first arrival P or PKP to escape detection if the later phase is observed. The amplitudes of later phases are generally smaller than the initial arrival. For this reason the single array phase identifier works by identifying a pair of signals if their parameters satisfy a certain relation.

With two arrays available, the object is to find earthquakes which have phases observed at both arrays. If the earthquake is large enough and well located so that both arrays receive at least just one phase and not necessarily the same phase type, it will be shown that it is relatively easy to identify the two phases and locate the earthquake. On the other hand if the earthquake is so small that one array misses it entirely the situation is almost identical to the single array case--the difference is that one will know not to expect the event to be observed at the other array. As was seen in the earlier chapter as of the time of this analysis, NORSAR indeed had poorer detection capability than LASA. The two array phase identifier was designed with the hope that both arrays would have equal capabilities and that there would be a substantial number of events common to both arrays. It was expected that with the two array phase identifier there would be detections at both arrays which would never be associated with an earthquake unless information from both arrays was available to the phase identifier.

The purpose of this chapter is mainly to set up the mathematical formalism of solving the identification problem. The next section is a brief review of the basic concepts of statistical pattern recognition and decision making and may be skipped with little loss of continuity. The final section of this chapter ties these concepts to the identification problem.

3.2 Pattern Recognition

Pattern recognition methods must perform two basic functions, (1) the characterization of a set of common pattern inputs that belong to the same class and (2) the classification of any input as a member of one of several classes.

For our purposes it shall be assumed that the observations of a specific pattern can be described adequately by a finite dimensional vector $\bar{X}$ which we shall term the _observation vector_. Thus, a pattern corresponds to a point in n-dimensional space. (For the two array phase identifier, $\bar{X}$ consists of the beam numbers of the detections from the LASA and NORSAR respectively and the time interval between their arrivals.) The object is to classify $\bar{X}$ into one of m categories and to have some estimate of the probability of correct classification. This basically partitions the observation space into m disjoint regions. The regions may not be simply connected.

The next assumption is the existence of a transformation on the $\bar{X}$ space that will cause points in the same class to cluster together. Hopefully this transformation will keep points of different classes in separated clusters. Except for certain special cases, there is no specific routine that will find the best transformation. If the clusters are adequately separated the proximity of a specific point to a cluster center should be a measure of how much certainty a point can be associated with a specific class. A more detailed discussion of this model is given in Sebestyen (1962).

The number and nature of the different types of classes may or may not be known. If the class types are unknown, cluster analysis methods could be used. In the identification problem dealt with here we are fortunate to have the different types of classes well defined.

The distinguishing characteristics of these classes or features may or may not be known. In our case, they are known partially. To extract these features a training set, i.e., a set of patterns with known classification, is used. Here, the classification of the elements in the training set are not known to complete certainty.

Patterns belong to the same class if they are similar or equivalent under certain operators. The measurement of their similarity requires the introduction of a metric.

In this dissertation, the development of the phase identifier will rely heavily upon statistical pattern recognition techniques. This method does not place any particular restriction on the nature of the clustering of patterns. Also this approach is very reasonable due to the probabilistic nature of the signal, the noise and the measurement errors.

The probabilistic model is used to describe the cluster distributions. Given a specific classification one can ascribe a certain probability that the observation coordinates fall at a certain point. This probability will reflect the degree of clustering of other pattern samples from the same class around the point.

The basic rule used in classifying the detections by the phase identifier is Bayes Rule. This rule will minimize the cost of making the wrong decision (Van Trees, 1968).

Suppose we have two sources generating an observable output $\bar{r} = (r_1, r_2 \dots r_n)$, where r_1, r_2 ... r_n consist of the observation parameters of the detection such as beam number, time of detection and intensity. The two sources generate a particular point in observation space with conditional probability densities $p(\bar{R}|H_0)$ and $p(\bar{R}|H_1)$ where $p(\bar{R}|H_i)$ means the probability of output $\bar{r} = \bar{R}$, given the hypothesis H_i that source i (i = 0 or 1) generated $\bar{r}$. The sources are hidden in a black box so that it is impossible to tell which source generated the output. In our problem these two hypotheses could be:

H_0: detection due to noise (false alarm)
H_1: detection due to seismic event

The discussion here is confined to decision rules that are required to make the choice. Each time the experiment is conducted one of four things can happen:

(1) H_0 true and H_0 chosen
(2) H_0 true but H_1 chosen
(3) H_1 true but H_0 chosen
(4) H_1 true and H_1 chosen.

The Bayes Rule makes the following assumption. The first is that the probability that source i generated the output is known and is denoted P_i, the a priori probability. The second assumption is that a cost C_{ij} is assigned to each possible action. C_{ij} is the cost of choosing hypothesis i when actually hypothesis j is correct. Thus, each time an experiment is done a certain cost will be incurred. It is also assumed that the cost of making the wrong decision is greater than the cost of a correct decision. It is known (Van Trees, 1969) that the decision criterion that will minimize the loss on the average is Bayes Rule. The decision rule is basically the following: Compute the ratios

$$A_0 = p(H_0|\bar{R})(C_{10}-C_{00}) \quad A_1 = p(H_1|\bar{R})(C_{01}-C_{11}) \tag{3.1}$$

using

$$p(H_i|\bar{R}) = \frac{p(\bar{R}|H_i)P_i}{p(\bar{R})} \tag{3.2}$$

and choose the hypothesis with the largest A_i. In many cases the cost matrix is unknown. The test then maximizes $p(H_i|\bar{R})$, the <u>a posteriori</u> probability, and is called the <u>maximum a posteriori test</u> (MAP). If the a priori probabilities, P_i, are unknown, then the test maximizes $p(\bar{R}|H_i)$, the likelihood of $\bar{R}$ given H_i, and is called the <u>maximum likelihood test</u> (ML). ($p(\bar{R})$ is independent of the two hypotheses so it does not enter in the decision making). These tests can be easily generalized to more than two hypotheses. An equivalent formulation of the maximum likelihood test is the <u>likelihood ratio test</u>. In this test one evaluates

$$\Lambda = \frac{p(\bar{R}|H_1)}{p(\bar{R}|H_0)} \tag{3.3}$$

If Λ is greater than a threshold T (T=1 if $P_1=P_0$), H_1 is accepted. Otherwise ($\Lambda < T$) H_0 is accepted.

As a simple example, consider the following particular case. Suppose the probabilities of observing $\bar{r}$ under the two hypotheses are both Gaussian with zero means but different variances σ_1^2 and σ_2^2. The experiment consists of making N separate observations, $r_1, r_2 \ldots r_N$.

Thus

$$p(\bar{R}|H_1) = \mathop{}\limits_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_1} \exp(-R_i^2/2\sigma_1^2)$$

and (3.4)

$$p(\bar{R}|H_2) = \mathop{}\limits_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_2} \exp(-R_i^2/2\sigma_2^2)$$

The logarithm of the likelihood ratio is

$$\ln \Lambda = \frac{1}{2}\left(\frac{1}{\sigma_2^2} - \frac{1}{\sigma_1^2}\right) \sum R_i^2 + N \ln \frac{\sigma_2}{\sigma_1} \qquad (3.5)$$

The Bayes Rule is to select hypothesis H_1 if $\ln \Lambda > \ln T$ and otherwise H_0. (Since the natural log function is monotonic increasing, the inequality is not destroyed by taking logarithms). The only unknown quantity in this test is $\sum_{i=1}^{N} R_i^2$, which shall be denoted as $l(\bar{R})$. The test can be rewritten as

$$l(\bar{R}) \underset{H_1}{\overset{H_2}{\gtrless}} \frac{2\sigma_1^2\sigma_2^2}{\sigma_2^2-\sigma_1^2} \left(N \ln \frac{\sigma_1}{\sigma_2} - \ln \eta\right) \qquad (3.6)$$

if $\sigma_2 > \sigma_1$. The main point to be drawn from this example is that the decision is based upon a scalar quantity $l(\bar{R})$. A second important point is that $l(\bar{R})$ is basically a measure of the distance of the observation vector from the origin. This will be seen again.

In the classification problem on hand both the cost matrix C_{ij} and the a priori probabilities are not known. These variables determine the <u>threshold</u> term T. The threshold controls the relative number of the two types of errors. If T is set high then H_0 will be selected more frequently. There will be more errors of the type where H_0 is chosen while H_1 is true and fewer errors of the other type. For example, if hypothesis H_1 is signal and H_0 is false alarm, this would mean that more false alarms would be mistaken for signals. Lowering T will have the reverse effect.

Usually the decision threshold parameter is unspecified variable since the costs and a priori probabilities are merely educated guesses. The relative number of the two types of errors is estimated either from theory or experiment as a function of T and the most practical value is used.

The performance of the decision processor is a measure of how often the right decision is made. The performance depends on how well the observation parameters separate the two hypotheses. In other words, it depends on the dissimilarity of the output from the two sources. In the given example the performance is determined by the ratio of the two variances σ_2^2 and σ_1^2. If $\sigma_2^2 = \sigma_1^2$, the two hypotheses become degenerate with respect to the observation parameters.

3.3 The Training Set

In the last section the formalism for the classification problem was discussed and the fundamental principles of statistical decision theory were reviewed. This section consists of a short interface to the next two chapters, where the single array phase identifier and two array phase identifier are described and evaluated. For both of these identifiers it was necessary to transform the detection parameters to another set of coordinates so that the different classification of detections would cluster in the observation space. Since the transformation is essentially the same in both identifiers, it is appropriate to discuss it in this chapter.

For each single detection in the LASA and NORSAR log, the detection processor records the exact time the strongest beam goes into detection state, the beam number, the total time duration of the detection state, the Maximum Short Term Average (MSTA), and the Long Term Average (LTA) just before the detection state. The MSTA is the largest STA value while the beam is in detection. As described in Section 2.2, STA for LASA is the mean of 0.8 seconds of the digitized, filtered, rectified clipped, beam data sampled at 20 Hz. STA is measured in so-called quantum units where 1 quantum unit is set at nominally 0.028 millimicrons for LASA; NORSAR has a different digitization level.

The MSTA should be reflective of the amplitude of the signal. Since the incoming signal is not usually perfectly in line with a beam, and since it is not the peak signal but a 0.8 second average near the peak, MSTA will tend to underestimate the actual signal amplitude. The analyst determination of the signal amplitude is fortunately reported in the summary bulletins. Matching these reports to the largest detections in a signal group, we calibrated the MSTA measurements independently. The matching criterion is discussed in Appendix A. Figure 15 shows the maximum MSTA in a detection group versus the quoted amplitude. In order to reduce the scatter substantially in the plots the data points were averaged whenever possible over 0.1 mμ units. One of the reasons for the larger scatter is that NORSAR signal extends through a higher frequency band. (The analyst measures the peak of the signal in the first few seconds). Another reason is likely poorly directed beams. It is expected that scatter for the NORSAR data will eventually be reduced very considerably.

In developing the phase identifiers only the start time of the detection, beam number and MSTA were used as input. It was not felt that LTA, signal duration or number of detections in a group introduces any substantial additional information. In many instances the LTA becomes contaminated by the earlier part of the signal. The LTA is correlated with the MSTA for the moderate size signals. The signal duration and number of detections were also correlated to the MSTA's. For this reason, it was believed that an insignificant amount of additional information would be introduced if those parameters were included at the expense of more computational time.

Vast amounts of computer memory and training data would be needed if the beam numbers were not transformed into more suitable coordinates. The beam numbers give very little indication of the direction of a beam and even more important, how close one beam is to another. For this reason it was desirable to convert these beam numbers to a more physical quantity. There was a choice of using either the velocity azimuth coordinates of the beam or the geographic coordinates of the beam assuming a phase interpretation. The latter was used in both phase identifiers for the following two reasons: (1) To compute travel times of the phases it would always be necessary to convert to the geographic coordinates; and (2) the actual velocity, azimuth of the beams is probably partially affected by the type of phase.

Though both LASA and NORSAR list the beam coordinates in the detection log, it was still necessary to make our own calibration. The reliability of these coordinates was uncertain and in addition the figures were listed only for the P and PKP waves. The calibration was done using the training set. This set was generated using ERL epicenter determinations. The arrival times of the various body wave phases were predicted for NORSAR and LASA and were matched to their respective detections in the log. About 2200 matches were made. Many of the predicted later phases could not be matched to any of the detections. (The matching criterion is discussed in Appendix A). Most of the detections in the training set belonged to the LASA detection log. Due to NORSAR's inferior detection capability, fewer later phases were detected by NORSAR.

The training set was sorted into the array, phase type and beam number. For the seismic beams there were generally many identified detections. However, 210 of the 600 LASA beams and 220 of the 510 NORSAR beams had no training events at all.

As discussed in Appendix A any matching criterion will accept a certain number of false matches. A predicted arrival could be matched to a false alarm detection or a phase of a different earthquake which happened to arrive at almost the same time. A wrong beam could be triggered by signal looking through the sidelobe of that beam.

Using the nominal beam positions listed by SAAC, the false matches were removed subjectively. Generally, it was expected that the quoted azimuth of the beam and the azimuth determined from the ERL epicenter to be within 30 degrees of each other. However, if the signal came in strong enough to preclude the possibility of a false alarm and at almost the predicted time of the detection, then this restriction was relaxed. If the beam had many matched detections, then it was fairly easy to spot the bad matches, since the location of the event for those matches would be completely off. There was a considerable number of cases where it was very difficult to decide whether to accept the match. For example, the PcP phase comes in within 60 seconds of the P arrival for earthquakes at distances greater than 55 degrees from the array. In these cases it was sometimes very uncertain whether the PcP phase was correctly matched, or it was matched to either a depth phase, aftershock, or part of the coda. The resolution of the beam was sometimes not sufficient to distinguish the phase velocities of the P and PcP which gradually approach each other. Usually a PcP match was rejected if the distance of the training event was almost the same as the P training event. Often the same beam would be triggered about 30 seconds later and be matched to a PcP. The second difficult case was the SKP phase which arrives 205 seconds after the PKP phase in the same beam. The SKP similarly could be confused with a depth phase, coda or aftershock.

For the above reasons the generation of the training set involved a considerable amount of subjectiveness. Since the performance of the phase identifier could mainly be evaluated only on the basis of the training set, there was a considerable amount of laxness in testing the phase identifiers. If another set of detection log data was available with the beams steered exactly the same way, then it would have been possible to make a more objective evaluation. Unfortunately, errors were found very recently in both the LASA and NORSAR beam station corrections. The implementation of the new station corrections may require recalibrating the beams with another training set.

On the basis of the training set, a table transforming the beam numbers to geographic coordinates was made. The coordinates, of course, depended on the phase type. (Some beams could detect as many as six different phases). This table was referred to by either the single array or two-array phase identifier.

3.4 Summary

Chapter 3 laid the groundwork for both the single and two-array phase identifiers. The classification problem was divided into one of separating the false alarms from signals, and of distinguishing the different types of body wave phases P, PKP, PcP, ScP, SKP, PP, PKKP and P'P'.

Chapter 4

Classification of Detections Using One Array

4.1 Introduction

In the last chapter the theory of the classification of detections was described and the training set was created and used to calibrate the beams. In this chapter we apply the previous results to the single array problem.

4.2 Single-Array False Alarm Discrimination

The LASA detection logs list, on the average, 500 detections a day. Many of the weak detections are questionable. The strong detections reflect the world seismicity pattern, however, the weak detections are uniformly distributed among the 600 beams. The similar effect is also observed for NORSAR detection logs. Because there is no evidence to believe that there are many low magnitude earthquakes occurring uniformly over the various aseismic and seismic regions of the world, it is believed that the weak detections are not real signals. Hence they are called false alarms.

It is very difficult to identify a specific detection as a false alarm. All existing earthquake catalogues generated today are only reporting a fraction of the actual occurring earthquakes. The false alarms are generally due to a sudden increase in microseismic noise which is enough to trigger one of the beams.

The goals of this section are (1) to find a criterion for distinguishing signals from false alarms, and (2) to estimate the number of detections that are real seismic signals, and the number that are false alarms.

To estimate the number of false alarms and signals at LASA and NORSAR, a statistical study was performed. Seismic and aseismic beams were distinguished by counting the number of detections per beam above a certain MSTA threshold. (MSTA is the Maximum Short Term Average, as defined in 3.3). The threshold was chosen to exclude most of the false alarms. Next a set of aseismic beams with no detections above that threshold was found. The distribution of MSTA for detections from these aseismic beams was determined (49 aseismic beams were used for LASA and 27 aseismic beams for NORSAR). This was assumed to be the distribution of MSTA for false alarms. (It is possible that a few real signals may have contaminated the false alarm distribution, due to leakage through the sidelobes of beams, but the effect is negligible). The false alarm distribution was then extrapolated to all 600 beams assuming that they occurred uniformly.

The total MSTA distribution of all detections in the log was also determined for the same time period. This total distribution included both signal and false alarms. The difference between the total distribution and extrapolated false alarm distribution would reflect the MSTA of the signals. In Figure 16 the extrapolated false alarm and total MSTA distributions were plotted for LASA and NORSAR. (The false alarm distribution exceeds the total distribution at low MSTA's due to the magnification of statistical error in the extrapolation of the false alarm distribution). It is evident from the figure that the false alarms dominate the distribution for the weakest detections but become a smaller fraction of the detections as MSTA increases. This is as it would be expected, since background noise is generally small and relatively constant.

The actual MSTA distributions reflect very many factors. The distribution goes down with MSTA since the frequency of large earthquakes goes down with magnitude according to Richter's relation (1958). It is too complicated to explain the distribution of MSTA's analytically, since it largely depends on the Detection Processor algorithms and signal waveform. There are usually several detections with different MSTA's reported within a few seconds of each other for the same seismic signal.

The probability of a detection being a false alarm, given the MSTA, was determined from the previous distributions and was approximated by a straight line for the range of interest. Figure 17 plots the probability of LASA and NORSAR detections being false alarms.

The MSTA is the strongest criterion of distinguishing signal from false alarm. If MSTA for a LASA detection is above 350, then the possibility of a detection being a false alarm is ruled out completely. If the LASA MSTA is below 100 then it is more likely a false alarm than a signal. MSTA values for LASA signals range up to several thousand, so the false alarm region is small in comparison to the possible range of the parameter. Unfortunately, very many signals have strengths in the false alarm range.

The optimum signal-false alarm discriminator based on the detection log data would probably use seismicity information in addition to MSTA. The ratio of signal detections to false alarm detections depends very strongly on the beam number. This ratio varies over a range of .70 to nearly 0, depending on whether the beam is pointed at a very seismic area or a completely dead area. Furthermore, the signal-false alarm discriminator could also use the fact that earthquakes tend to cluster in time and space due to the existence of aftershocks (Shlien and Toksöz,1970b), while false alarms have very little of this tendency. For example, if 10 detections have been reported by the same beam within a period of two days, at least 8 of these detections are likely to be real signals. (Less than one false alarm is detected at LASA per beam per day). Consideration of these observations would, of course, improve the performance of the discriminator. It would also have the effect of biasing the discriminator against earthquakes in aseismic regions, which do occur occasionally. A more mathematical discussion on the discriminator has been put in Appendix C.

Signals listed in the summary bulletins are seismic phases which the analyst believes he definitely sees in the properly steered beam trace. (They may be so small that they would be invisible in any subarray trace). Assuming that the seismic phases were real, they were matched to the biggest detection in the detection log, and the MSTA distributions for these matched detections were determined. In Figure 18 we plot the empirical cumulative probability distribution function of the MSTA of these matched detections for LASA and NORSAR. From these distributions one can read off the number of signals reported by SAAC that would be missed if the signal-false alarm discriminator removed all detections below a certain MSTA. Though the fraction of signals deleted are very substantial past the false alarm region, it should be noted that these signals are very small events ($m_b \simeq 3.5$ for LASA).

Not all signals detected by LASA and NORSAR are reported in their respective summary bulletins. LASA for example will not report any event beyond 100 degrees even if it is very visible. Besides, there are probably real signals which the Detection Processor flags but which the analyst ignores because he cannot see them on the beam. The number of signals detected by LASA and NORSAR was estimated as a function of MSTA by counting the number of detection groups in the respective logs and subtracting off the estimated number of false alarms. A detection group was defined as a set of detections occurring within 30 seconds of each other. The number of false alarms was estimated from the distributions in Figure 17. (False alarms may also come in groups). In Figure 19 we plot the cumulative number of signals versus the maximum MSTA of the detection groups for LASA and NORSAR. Also plotted for comparison is the cumulative number of signals reported in the summary bulletins versus the maximum MSTA of the matched detection groups. LASA apparently detects more than 60 signals a day and NORSAR more than 25. An independent study being performed by the Seismic Discrimination Group at Lincoln Laboratories confirms this fact. About 60 earthquakes a day could be verified by looking at the seismograms of neighboring stations (Russell Needham, personal communication). Of course, if LASA alone attempts to detect all these events, it will also have to accept very many false alarms. Figure 19 also shows the cumulative number of false alarms that would have to be accepted if the signal-false alarm discrimination accepts anything above a certain MSTA threshold.

From these figures it is again apparent that NORSAR does not have the same detection capability as LASA. LASA detects twice as many signals than NORSAR. Furthermore, the false alarm problems seems more severe for NORSAR. When LASA detects 25 signals a day, the false alarm rate for the same signal rate is 11 per day. The false alarm rate is determined by the noise level at the array. Part of this discrepancy can also be attributed to the different seismicity distribution around NORSAR. There are considerably fewer earthquakes occurring within the 20-90 degree distance range from NORSAR than LASA. For example, the South American Seismic belt is beyond the shadow zone from NORSAR, but within 80 degrees of LASA.

Due to the problem that a signal or false alarm may trigger several detections we had to estimate the number of signals that LASA or NORSAR detects in a rather roundabout fashion. Basically, the ratio of false alarms to signals SF(MSTA) was estimated as a function of MSTA using seismic and aseismic beams. Next, the number of detection groups, DG(MSTA), was determined as a function of MSTA. The cumulative number of signals CSIG(MSTA) and false alarms CFA(MSTA) as a function of MSTA was computed essentially from

$$\mathrm{CSIG(MSTA)} = \int_{\mathrm{MSTA}}^{\infty} \mathrm{(SF)(DG)\ dMSTA} \tag{4.1}$$

$$\mathrm{CFA(MSTA)} = \int_{\mathrm{MSTA}}^{\infty} \mathrm{(1-SF)(DG)\ dMSTA} \tag{4.2}$$

4.3 The Single Array Phase Identifier

The last section dealt with the problem of distinguishing signal from noise. In this section we examine the problem of classifying the signals into the different phase types. These two problems were treated separately for convenience. The approach to this problem is considerably different since it is necessary to rely on contextual information. Later phases cannot be identified using a single station unless they can be related to the first arrival (P or PKP).

The input to the phase identifier is a pair of detections which have occurred within 30 minutes of each other. (No attempt has been made to find later arrivals after P'P'). The parameters of the pair of detections are tested with respect to eleven different hypotheses listed below.

Hypothesis	First Detection	Second Detection
1	P	PCP
2	P	SCP
3	P	PP
4	P	PKP
5	P	PKKP
6	P	P'P'
7	PKP	PP
8	PKP	SKP
9	PKP	PKKP
10	PKP	P'P'
11	none of the above	

The last hypothesis includes PKKKP, SKKP phase, other combinations of these phases such as PcP - ScP, phases of distinct earthquakes, and the possibility of one or both of the detections being false alarms.

The phase identifier is based on the following fact. For many of the first 10 hypotheses the phase velocities, azimuths and time difference of the two detections bear a certain relationship with each other depending on the distance of the earthquake and the hypothesis. For example, both signals either arrive in the same azimuth or in exactly the opposite azimuth. Both PKKP and P'P' phases travel more than halfway around the earth and arrive at the station from the back azimuth. Further, given the distance of the earthquake, then the specific phase will arrive at certain times and with certain velocities. In Figure 20 the travel time interval between first arrival and later phase is plotted vs. the inverse phase velocity of the later phase. The inverse phase velocity of the first detection could be plotted on an axis coming out of the paper. Thus, the curves for the different hypotheses are actually separated in three-dimensional space. If the parameters of a detection pair lie remote to any of these space curves, then the phase identifier would choose hypothesis 11. On the other hand, if parameters of the detection pair lie near a specific phase curve like P-P'P' then either it happened to be a coincidence or else the two detections are actually P and P'P' respectively. Since the probability of a coincidence is small, the second hypothesis is more likely.

This picture expresses the basic principle of phase identification. The picture is similar for the two-array phase identifier, but in the two-array case the azimuths and inverse phase velocity of the LASA and NORSAR detections are coupled to each other by the spherical geometry. (This is discussed in further detail in Chapter 5).

The actual implementation of the single array phase identifier is quite different for practical considerations, but the basic principles are the same. For each detection pair, the phase identifier tries each of the phase interpretation hypotheses. The best interpretation is chosen using statistical techniques. The input parameters used for every detection pair are the beam numbers of the former and latter detections, NBM1 and NBM2, the Maximum Short Term Average of the two detections MSTA1 and MSTA2, and the time difference between the detections, ΔT. The likelihood ratios of hypotheses 1-10, over hypothesis 11 are each computed as follows:

$$\Lambda_i = \frac{p(\mathrm{NBM1},\mathrm{NBM2},\mathrm{MSTA1},\mathrm{MSTA2},\Delta T \mid H_i,E_i)}{p(\mathrm{NBM1},\mathrm{NBM2},\mathrm{MSTA1},\mathrm{MSTA2},\Delta T \mid H_{11},E_i)} \tag{4.3}$$

$p(\mathrm{NBM1},\mathrm{NBM2},\ldots,\Delta T \mid H_i,E_i)$ is the probability (likelihood) of having two detections ΔT seconds apart with parameters NBM1, NBM2... given that the detections are interpreted by hypothesis i and the location of epicenter is E_i. Note that E_i is a function $E_i(\mathrm{NBM},H_i)$ of both the beam number and hypothesis. The identification of detections is always involved with the location of the epicenter. The input parameters and phase interpretation specify almost the earthquake's coordinates.

The a priori probabilities of the different hypotheses were found generally to be within less than an order of magnitude of each other on the basis of the training set. Very little would be gained by including them in the test. For this reason the a posteriori probabilities were not computed.

Let us now describe the estimation of Λ_i and the performance of the phase identifier. It is very awkward to estimate Λ_i from the original input parameters, since the parameters are not mutually independent. If the original parameters could be transformed to a new set $S_{1i}, S_{2i}, S_{3i}, \ldots$ of independent parameters, then Λ_i could be evaluated simply as

$$\Lambda_i = p(S_{1i} | H_i, E_i) p(S_{2i} | H_i, E_i) p(S_{3i} | H_i, E_i) \ldots \qquad (4.4)$$

The following set of transformed parameters have that desirable property and are very convenient on the basis of programming considerations.

$$\begin{aligned} S_{1i} &= \mathrm{DIS(NBM1)} - \mathrm{DIS(NBM2} | H_i) \\ S_{2i} &= \mathrm{DIS(NBM1)} - \mathrm{DIS(}\Delta T | H_i) \\ S_{3i} &= \mathrm{AZ(NBM1)} - \mathrm{AZ(NBM2} | H_i) \\ \mathrm{MSTA1} &= \mathrm{MSTA1} \\ r &= \ln(\mathrm{MSTA2/MSTA1}) \end{aligned} \qquad (4.5)$$

where DIS is the distance of the epicenter from the array determined from either the beam number or travel time interval assuming hypothesis H_i, and AZ is the azimuth of the beams assuming H_i. E_i has been suppressed and H_i is kept only in the terms where it is actually used in the evaluation of S_i. Since the first detection is always tested as a P or PKP, depending only on the beam number, the interpretation H_i only affects the second detection. The parameters S_1, S_2 and S_3 also have the valuable property that for a correct identification the distances or azimuths of the two terms will match and the parameters will be close to zero.

The probability distribution functions were easily evaluated from the LASA training set. Though the probability distribution functions do depend on the hypothesis H and the distance of the earthquake, the differences between hypotheses H_i $(i=1,2,\ldots 10)$ are small enough to warrant neglecting them except for H_{11}. Subscript i has been suppressed on S_1, S_2 and S_3. (A small compensation was made in the actual program for distance by scaling parameters S_1 and S_2 for events near the shadow zone).

For H_{11}, the complement of all the former hypotheses, the distribution of parameters $S_1, S_2, \ldots$ could not be estimated from the training set of detections. A new detection log was generated with the same statistical properties of the former log except that the detections were unrelated to each other. This was done by shuffling the original detection log by the method described in Appendix D.

In Figure 21 the distributions of the parameters S_1, S_2, S_3 determined from the LASA training set and the shuffled detection log are shown. The differences between the two columns imply the feasibility of distinguishing H_{11} from all the other hypotheses.

The parameters S_1, S_2, and S_3 were found to be uncorrelated near the origin. The correlation matrix determined from 236 training samples was

$$\begin{array}{cccc} & S_1 & S_2 & S_3 \\ S_1 & 1 & .23 & .35 \\ S_2 & .23 & 1 & .13 \\ S_3 & .35 & .13 & 1 \end{array}$$

Most of the correlation was found when the S_i parameters take extreme values. In this correlation determination we excluded S_i with absolute value greater than 8.

Parameter r is a measure of the relative attenuation of the later arrival with respect to the first arrival. Except for phases coming in from the shadow zone the later phase is nearly always attenuated with respect to the first arrival. The amount of attenuation does depend on H_i. The P'P' goes through the earth's core twice so that it is much more attenuated than, say, PcP. PP, ScP and SKP tend to have more energy in the longer periods. The frequency response of the filter in the detection processor tends to attenuate the lower frequencies. The P or PKP phase becomes unusually attenuated at the shadow zone. The amount of attenuation becomes very comparable to that of PP, PKKP and SKP so that the later phase sometimes comes in stronger than the first arrival. In Figure 22 r is plotted using the training set for the different interpretations. Large scatter is due to the inherent variability of the amplitude data. The above mentioned effects are still very apparent.

Normal approximations were made to most of the above parameters. The means and variances of the distributions were determined by plotting the cumulative distribution on normal probability paper. This way the effect of extreme data points could be minimized. The MSTA1 distribution, however, could not be approximated by a normal distribution. The means and variances of S_1, S_2, S_3 and their normal approximations used to estimate $p(S_1)$, $p(S_2)$, etc., are listed in Appendix E. Further details on how the MSTA distributions were approximated are also included in the same appendix.

The single array phase identifier programmed in basic Fortran was tested on 100 days of data. The program works as follows. All detections which have occurred within the last half hour are stored in a memory buffer. A new detection is read off from a magnetic tape and then tested with respect to each detection in the buffer. For each pair of detections the log likelihood ratio l_i is evaluated for the different hypotheses $i = 1,10$. If l_i is less than a chosen threshold T then this implies that the ith hypothesis is probably wrong and the phase identifier goes on to the next hypothesis (Selection of T is described later). If all the hypotheses are rejected, then that detection pair is forgotten. On the other hand, if l_i is above the threshold, then that hypothesis becomes a reasonable prospect. The other hypotheses are still tested and the one which has the largest l_i is accepted by the phase identifier. The phase identifier prints out the detection parameters of the detection pair, the phase identification of the two detections, the log likelihood ratio statistic, and the earthquake's epicenter and origin which generated the observed signals.

Many shortcuts are taken to expedite the execution of the phase identification. The transformations from beam numbers and travel time interval to distance and azimuths (4.1) for a given detection pair and hypothesis is done using a table look-up. Interpretations are rejected outright if the time interval between detections is outside its expected range. If a particular phase is never observed by that beam in the training set, then the hypothesis is categorically ignored. The consequences of this procedure are really not so bad as one may think. Since a phase usually triggers several beam detections the probability of accidentlly rejecting that phase interpretation is low. Therefore only a few later phases coming from aseismic regions would be missed.

Execution time using the Lincoln Laboratory PDP-7 was very short. One hundred days of the LASA detection log were processed in two hours. With a threshold level T = 0, an average of 9.5 later phases were found per day in the detection log. About 200 later phases were due to the aftershock sequences in the New Ireland and New Britain regions 14 July to 2 August 1971. Of these 9.5 later phases, 2.0 could be confirmed using the training set. Thus, the phase identifier found 7.5 later phases a day that did not exist in the training set. 2.3 of these 7.5 phases could be confirmed indirectly using the LASA summary bulletin. This leaves a total of 5.2 later phases that could not be checked by any simple means.

It would be expected that the phase identifier would occasionally pick out fictitious later phases due to situations where false alarms or independent signals fortuitously triggered the beams in the right sequence and times. Estimation of the number of fictitious phases that were found was done using the shuffled LASA detection log described in Appendix D. (It is too cumbersome to obtain theoretical estimates). Over a period of 20 days of the shuffled detection log, 45 fictitious later phases were found. Thus, 2.3 of the 7.0 later phases a day are probably due to unrelated detections occurring at just the right times.

If the threshold level T is raised the number of fictitious later phases picked can be reduced very considerably without missing too many real phases. The log likelihood ratio statistic, l_i, was above 4.0 for all later phases in the training set. None of the training later phases would be missed by the phase identifier with T = 4, but half of the fictitious later phases would be eliminated. With the threshold level T set to 4, 7.7 later phases a day were found.

On the basis of the training set, almost no misidentifications were made by the single array phase identifier. The confusion matrix for the 100-day trial run is shown in Table 1. The left column of the table lists the correct classification of the training phases. The top row of the table lists the identifications made by the program. Thus, the numbers along any row show the distribution of the phase identifier's classification of a set of particular training phases. There were almost no numbers off the main diagonal of the matrix. None of the training phases were classified as incorrect phase.

Evaluation of the phase identifier on the basis of the training set tends to make the performance appear much better than it is in reality. Since the table transforming beam numbers to distance and azimuth given a phase interpretation was determined from the training events (described in Section 3.3) the identifier is definitely biased towards picking out the training phases. Furthermore, the method of generating the training set would tend to delete any training detections that triggered the wrong beam. As described in Section 3.3, there was a considerable subjectiveness in deciding whether a phase predicted by ERL epicenter determinations was properly matched to the right signal.

If a less biased method of evaluating the phase identifier was possible we would not resort to the training set. Unfortunately, no set of pre-identified detections could be found or generated other than the training set. The LASA summary bulletins have stopped reporting later phases since January, 1971. Besides, the phase identifications that are made by the LASA analyst are also subject to error. Instances are known where SAAC misidentifies a PP phase for a P phase and reports an earthquake which has never occurred.

The single array phase identifier has one drawback. The basic assumption of the identifier was that the first arrival of an earthquake must be detected by LASA if a later phase is observed. This assumption is not always true for events arriving from the shadow zone. Many cases were found in an earlier study one year ago where the PP or PKKP phase is detected by LASA but the P or PKP phase failed to trigger the Detection Processor. As a result, the PP or PKKP phase is either not identified or is misidentified.

Another unavoidable source of error is the occurence of several earthquakes along the same azimuth within a time interval of a half hour. The azimuth of the detections is the most important decision parameter of the single array phase identifier. Distance errors and amplitude variations are large so that they tend to be secondary decision factors. This was apparent from Figures 21 and 22. (The inclusion of these factors, of course enhance the performance of the phase identifier). If two earthquakes do occur within the same azimuth and at the right times, then it may be in certain circumstances difficult to decide whether the two signals are independent or different phases of the same earthquake. It is possible that both of these hypotheses are correct, since distinct phases from two different earthquakes can easily arrive at one station at the same time. Luckily the occurrence of such coincidences are rare.

An analyst identifying the signals from the seismograms would have the same above two difficulties. He may nevertheless be able to use the wave shapes of the signals if they are strong enough to be seen.

4.4 Conclusions

The results of this study have shown that the automatic detection classifier is feasible, but it would still be desirable to have an analyst available who could refer back to the seismograms and check for any obvious errors. The automatic phase identifier would certainly save the analyst a considerable amount of time searching the seismograms or detection logs for later phases and testing and choosing the interpretations.

The automatic phase identifier runs at about 1000 times faster than real time. One day of detection log can be sifted through in a minute.

Later phases are a very small fraction of the signals detected by LASA. Since the presence of observable later phases requires fairly strong earthquakes, more than 90% of the seismic signals at LASA are P or PKP phases. LASA detects about 60 seismic signals a day. The estimated number of detected later phases is only 5 per day. When the LASA summary bulletin did report later phases, only 2 or 3 were reported per day. (During that period LASA bulletins did not report for more than 15 hours a day). Thus, later phases are likely to have little application in the confirmation of events.

If later phases are found they may nevertheless be used to improve the distance estimate of the earthquake. Arrival times of phases can be measured within one second. The travel time interval between phases is very sensitive to distance as seen in Figures 12 and 13.

False alarms were found distinguishable from signals, using the MSTA of the detection. False alarms are always weak detections. There is no way of distinguishing weak signals from false alarms using only the information in the detection logs. If one is willing to forego the signals smaller than 1 millimicron, then the false alarm problem is manageable.

The single array phase identifier was modified to run on

NORSAR's detection log. Only one later phase a day was found on the average. One fictitious later phase was found every 5 days on the shuffled NORSAR log. No extensive analysis was made.

Chapter 5

The Two Array Phase Identifier

5.1 Introduction

The last chapter described the classification of detections with one array. It was found possible to identify later phases automatically by matching them to their first arrival. With two separate arrays running simultaneously, detections from one array could be checked against another to find common events. If a matching detection is found by the other array it is unlikely that both detections were triggered by local noise. In addition, the epicenter parameters could be improved if data from two arrays are used. With two arrays, it is no longer necessary to find a first arrival in order to identify a later phase. For example, if LASA just observes PP and NORSAR detects just PKKP for the same event, then these phases can be identified unambiguously.

The extension of the single array phase identifier to two arrays involved very little new concepts. The fundamental principles adopted are exactly the same. For this reason the theoretical concepts described in the beginning of Section 4.3 will not be repeated here.

In the next section the design and testing of the two array phase identifier is described. The following section discusses briefly how one can improve the epicenter determination with information from two arrays.

5.2 Two-Array Phase Identifier

The description of the two-array phase identifier will closely parallel that of the single array phase identifier. The input to the two-array phase identifier is a pair of detections which have occurred within 30 minutes of each other. One detection is from LASA and the other is from NORSAR. The parameters of the two detections are tested against 50 different hypotheses. The first 49 hypotheses consists of all ordered pairs of the following phases: P and PKP, PCP, ScP, SKP, PP, PKKP and P'P'. (It was not necessary to distinguish the P and PKP phases, since the PKP phase is just the continuation of the P phase after the shadow zone). The hypothesis shall be labelled H_{ij}, where i is the phase at the first array and j the phase at the second array. The last hypothesis, $\bar{H}$, is similar to H_{11} of the previous chapter. It is the complement of the first 49 hypotheses.

The maximum likelihood ratio test is used to select the best hypothesis. Only three input parameters are used to make the decision. They are the LASA beam number, NBML, and the NORSAR beam number, NBMN, and the time difference between the two detections, ΔT. The likelihood ratios of hypotheses 1 to 49, over hypothesis $\bar{H}$ were computed as follows:

$$\Lambda_{ij} = \frac{p(\mathrm{NBML},\mathrm{NBMN},\Delta T \mid H_{ij},E)}{p(\mathrm{NBML},\mathrm{NBMN},\Delta T \mid \bar{H},E)} \tag{5.1}$$

where E is the presumed epicenter of the earthquake. Again the determination of epicenter coordinates is intimately related to the identification of the detections. The rest of the section described the estimation of Λ_{ij} and the performance of the two-array phase identifier.

The likelihood ratio test for the single and two-array phase identifier is not strictly optimum. The epicenter location, E, which is unknown, should be treated as an unwanted parameter in the identification process. The optimum test for the two-array identifier computes Λ_{ij}.

$$\Lambda_{ij} = \frac{\int p(\mathrm{NBML},\mathrm{NBMN},\Delta T \mid H_{ij},E)\,p(E)\,dE}{\int p(\mathrm{NBML},\mathrm{NBMN},\Delta T \mid \bar{H},E)\,p(E)\;dE} \tag{5.2}$$

where p(E) is the probability of the epicenter being at E. For practical reasons we did not try evaluating the two surface integrals.

The next step was transforming the beam numbers to a more convenient coordinate system so that Λ_{ij} may be evaluated. Distance and azimuth coordinates were preferred, since distance is needed to compute theoretical travel times of the phases. There was a choice of using coordinates centered at LASA or those centered at NORSAR. The calculations were duplicated in the two coordinate systems.

Beam numbers were converted to distances and azimuths from the respective arrays assuming a specific interpretation as follows.

$$\begin{aligned} D_L &= D(NBML|H_{ij}) \qquad & D_N &= D(NBMN|H_{ij}) \\ A_L &= A(NBML|H_{ij}) \qquad & A_N &= D(NBMN|H_{ij}) \end{aligned} \tag{5.3}$$

where

D_L is the distance corresponding to the LASA beam from LASA

A_L is the azimuth corresponding to the LASA beam from LASA

D_N is the distance corresponding to the NORSAR beam from NORSAR

A_N is the azimuth corresponding to the NORSAR beam from NORSAR

The above transformation only depends on the phase arriving at the particular array.

Spherical geometry was used to convert D_L, A_L, D_N and A_N to the coordinate system of the other array:

$$\begin{aligned} d_L &= d(D_N,A_N) \qquad & d_N &= d(D_L,A_L) \\ a_L &= a(D_N,A_N) \qquad & a_N &= a(D_L,A_L) \end{aligned} \tag{5.4}$$

where

d_L is the distance from LASA of the point specified by the NORSAR beam

a_L is the azimuth from LASA of the point specified by the NORSAR beam

d_N is the distance of the point from NORSAR specified by the LASA beam

a_N is the azimuth of the point from NORSAR specified by the LASA beam

The spherical transformation is given in Appendix F.

The theoretical travel time interval between the arrivals of the phases at the two arrays, Δt_X, was computed both ways, since one array has a better epicenter determination than the other, if they are not equal.

$$\begin{aligned} \Delta t_L &= \Delta t(\mathrm{NBML} | H_{ij}) \\ &= \Delta t(D_L, d_N | H_{ij}) \\ \Delta t_N &= \Delta t(\mathrm{NBMN} | H_{ij}) \\ &= \Delta t(D_N, d_L | H_{ij}) \end{aligned} \qquad (5.5)$$

The rest of the calculations almost mirrors the single array phase identifier. Parameters SA, SB, SC were determined from

$$\begin{aligned} \mathrm{SAL} &= D_L - d_L & \mathrm{SAN} &= D_N - d_N \\ \mathrm{SBL} &= A_L - a_L & \mathrm{SBN} &= A_N - a_N \\ \mathrm{SCL} &= \Delta T - \Delta t_L & \mathrm{SCN} &= \Delta T - \Delta t_N \end{aligned} \qquad (5.6)$$

These parameters again have the property that they tend towards zero for a correct identification and take on any value for a wrong identification. The distribution of these parameters was approximated by normal distributions as described in the Appendix G. The variances of these parameters depend on four factors, (1) the partition of the beam, (2) the inverse phase velocity and (3) its derivative with distance, and (4) the spherical geometry involved. Appendix G describes the details in estimating the variances. For $\bar{H}$, the complement of all other hypotheses, the distributions were again determined from a synthetic log described later. Λ_{ij} was computed from SAL, SBL and SCL, and again from SAN, SBN and SCN. The largest Λ_{ij} was used. We shall ignore the last identifier N or L in the above parameters. Hence SA, SB and SC.

The two-array phase identifier was programmed and tested on 89 days of data. The input was the LASA and NORSAR detection logs merged onto a single magnetic tape. The program works as follows: LASA and NORSAR detections occurring within the last half hours are stored in separate memory buffers. The program tries to match the current detection just read off from tape to a preceding detection in the memory buffer of the other array. The log likelihood ratio statistic l_{ij} is computed for the different hypotheses. If l_{ij} is less than zero the hypothesis is rejected and the next one is tested. If l_{ij} is greater than zero then l_{ij} is considered to be a prospect; the other hypotheses are still tested for the same detection pair. The hypothesis with the largest l_{ij} is accepted. The detection input parameters, the phase identifications, the log likelihood ratio statistic, the earthquake's epicenter and origin are all printed out. Similar shortcuts were made as described in Section 4.3 with similar consequences.

Execution time was about 5 times slower than the single array phase identifier. A total of 751 earthquakes were found to have phases common to LASA and NORSAR over a time span of 89 days (May to August, 1971). This corresponds to a rate of 8.4 events per day. 1.9 of the 8.4 events a day could be confirmed using the training set. This leaves 6.5 new events per day which were not wholly, or at all, in the training set. Of these 6.5 events, another 2.5 per day could be confirmed indirectly by either the LASA or NORSAR summary bulletins.

Some of the phase identifications and earthquakes found by the two-array phase identifier could be due to noise or independent signals fortuitously triggering the right LASA and NORSAR beams at the right times. These fictitious earthquakes cannot be identified, since no complete earthquake catalogue exists. The estimation of the number of such accidental occurrences is very cumbersome by theoretical methods since there are 49 different ways that a detection pair can be matched. The rate of occurrence of these false matches was determined using a synthetic detection log in which the LASA and NORSAR detections had the same statistical properties as before, except that they were completely independent of each other. Such a log was generated by merging the NORSAR log with the LASA log and incorporating an artificial two-day time lag in the NORSAR log. A total of 37 fictitious earthquakes were found in 35 days of the synthetic log. Therefore 1 out of 8.4 earthquakes a day found by the phase identifier is probably false.

8.4 earthquakes a day is very small in comparison to the total number of earthquakes LASA detects. It was shown in Section 4.2 that LASA detects about 60 seismic signals a day. NORSAR's detection capability at the time data was acquired was the biggest limiting factor.

The 8.4 events a day found by the phase identifier is a sizeable fraction of earthquakes reported in other bulletins. ERL reports 14 events a day, LASA Summary Bulletin reports 30 events a day and the NORSAR Summary Bulletin reports 6 events a day. (Since March, 1972, the number of events reported by NORSAR has almost doubled). The events found by the two-array phase identifier make up 40% of the ERL catalogue, 18% of the LASA Summary Bulletin and 60% of the NORSAR Summary Bulletin.

Using the training set an estimate was made of the number of phases that the two-array phase identifier classified correctly and incorrectly. They are listed for LASA and NORSAR for the different hypotheses in Table 2. On the basis of the training set the two-array phase identifier performed very well. As was discussed in Section 4.3, the evaluation of the phase identifier on the basis of the training set tends to make the performance look better than it is.

Epicenter determinations of the two-array phase identifier were within one or two degrees. The determinations are better than could be made with just the detection log of the LASA array. Epicenter determinations were of the same quality as the LASA Summary Bulletin and much better than the NORSAR Summary Bulletin prior to March, 1972. In the next section, we go into further detail on how the earthquake location is estimated and how it may be improved if one of the arrays detects additional phases from the earthquake.

It is more complicated to study the sources of errors with two arrays, since they are more dependent on the location of the earthquake. The two-array identifier does not have the same circular symmetry as the one-array identifier. The obvious sources of error are generally the same as for the single-array phase identifier. Phases having similar travel times and phase velocities are easiest to confuse. For example, the distinction between SKP and PKP phases becomes fairly fine, since they both arrive from beyond the shadow zone where distance determinations are inaccurate; they arrive within 200 seconds of each other; and they arrive often in the same beams. If both PKP and SKP are detected by one array, then the SKP could be identified fairly easily by the same array. Similarly, the two-array phase identifier may have difficulty distinguishing the PP from the SKP, and the P from the PcP at the distances where they both tend to arrive at similar times.

A different problem is identifying the P'P' (df branch) and PKKP (bc branch) phases. Both of these phases have high velocities so that they pass the array in the order of a second. As a result, the azimuth determinations may have a large error. Furthermore, since these two phases are not seen at close ranges, there is a resulting larger uncertainty in the epicenter's location. This, coupled with the fact that the phases come in very weakly, makes it very difficult to identify them.

In most of the cases P or PKP phase is involved in one or both of the matched detections. The later phases are generally only seen for the few large earthquakes. About 3 later phases a day at LASA could be matched to NORSAR detections. When both later phases and the first arrival can be matched to a phase at the other array, then the epicenter determination can be improved substantially. This will be illustrated in the next section.

5.3 Locating Earthquakes with Two Arrays

Accurate determination of an earthquake's epicenter largely depends on having many seismic stations distributed around the epicenter and knowing travel times of the phases exactly. Because a large-aperture seismic array is not particularly suited for precise determination of epicenters, the emphasis here has not been on the locating of events. Of course it would be desirable to be able get the best epicenter determination as possible with two arrays so that one does not have to wait as long for the data to be collected from all the other seismic stations.

Location of earthquakes with two arrays is better than with one. The object of this section is mainly to indicate non-mathematically what information is available, how it should be used, and what computational difficulties are to be anticipated.

To begin, we shall describe how the two-array phase identifier locates the epicenter in more detail. The time interval between the LASA and NORSAR detections and the interpretation of the detections defines the two finite non-intersecting curves on the surface of the earth. Any epicenter on those curves would satisfy the requirement that the predicted travel time interval of the two particular phases matches the observed time interval. The curves are fairly thin due to the small uncertainties in the measurements. If detections from a third seismic array were available, then another two locii of points would be defined satisfying the travel time interval between the other pair of arrays. The intersection of these locii would define the two possible epicenter locations compatible with the arrival times of the phase.

Since there are only two arrays, the ambiguity in location must be resolved using the beam locations. The width of the beams are usually much larger than the line defined by the travel time interval ΔT. On the basis of the two beam locations, which may or may not coincide and the ΔT

curve, one may determine the a posteriori probability density function of the epicenter. Maximizing this function with respect to the epicenter location will give the best location.

Though this is the optimum way of locating the earthquake with the two arrays, computationally this is very slow. The ΔT curve cannot be defined analytically, since it depends on the travel times of the phases which were determined empirically. The curve must be computed point by point and then interpolated so that one can compute the shortest distance of any prospective epicenter from the curve.

The two-array phase identifier uses a much faster method which does not give a location as accurately as above. The difference between the accuracy of the two methods is negligible. The method can be easiest explained by using Figure 23.

A magnitude 5.8 earthquake occurred in the Tonga Trench at 2:00 p.m. May 1, 1971. Several LASA and NORSAR beams were triggered by the events. The location of the LASA and NORSAR beams (as determined from the training set) are plotted in Figure 23 by L's and N's. The X is the actual epicenter. The continuous curve passing through the epicenter was determined by the time interval between the LASA and NORSAR first arrivals. The two-array phase identifier chooses one of the L's or N's as the epicenter. The beam chosen is the one closest to a beam of the opposite array and nearest to the ΔT line. This minimizes the log likelihood ratio statistic l_{ij}. The $\underline{L}$ was the epicenter presumed by the phase identifier.

Because of the large magnitude of the event, several additional phases were also detected. The ΔT curve based on the P - PP time interval and the P - PKKP time interval are plotted with dashed lines. The intersection of the three travel time interval curves lies much closer to the actual epicenter.

The accuracy of the later phase method is better than the conventional method used in the identifier. Finding the intersection of these curves involves numerical solution of nonlinear equations. The precision of this method depends on the geometry, and the derivatives of the travel time curves of the phases with distance. Clearly, it is most desirable to have the ΔT curves intersect with an angle close to 90 degrees.

In order to use this technique to its fullest capacity, two other effects must be taken into account. Due to various inhomogeneities in the earth such as dipping plates, the travel time tables could be off by as much as 5 seconds (Davies and McKenzie, 1969). With the use of later phases, these corrections could be determined at least relatively for the ray paths to LASA and NORSAR. The second correction has to be made for depth of earthquake. The depth is not known unless many later phases are observed. Then a set of nonlinear equations could be solved for epicenter, origin and depth together.

5.4 Conclusions

The identification of phases with two arrays is easier (though not less complex) than with one array. 50 different hypotheses could be distinguished with little error. About 8 earthquakes a day were found common to the LASA and NORSAR detection logs - one of them probably being fictitious. It is expected that this number will improve with NORSAR's new detection algorithm and station corrections. On the basis of the training set, the number of misidentifications were very small.

The two-array phase identifier is much more complicated and slower than the single-array identifier due to the more data that is analyzed and the additional computations in transforming from LASA coordinates to NORSAR coordinates and estimation of variance of parameters. However, it performs better than the single-array identifier in the problem of estimating the earthquake's epicenter. If the earthquake is large enough so that additional phases are found at either array besides the original pair, then the epicenter location estimate may be improved very substantially using travel time interval curves.

Chapter 6

Conclusions

In this study, the capabilities of LASA and NORSAR were evaluated on the basis of their present signal processors. The statistical properties of the output of their Detection Processors were determined. The problems of discrimination of signals from false alarms, identifying later phases with one and two arrays, and the determination of epicenter location with two arrays were investigated. The results of this analysis are listed here.

(1) LASA detected over 80% of the ERL epicenter determinations in the distance range 20 to 90 degrees from LASA and over 75% of these epicenters beyond 80 degrees.

(2) NORSAR detected (at the time of the analysis) about 60% of the ERL events between 20 and 80 degrees from NORSAR and about 35% of the events beyond 80 degrees.

(3) The LASA Summary Bulletin reports 3 times as many earthquakes as ERL in the distance range 10 to 95 degrees. The LASA seismicity distribution faithfully mirrors the ERL seismicity distribution in the above range.

(4) LASA and NORSAR body wave magnitude determinations do not show any easily detectable biases with respect to the ERL magnitude determinations.

(5) On the basis of the frequency-magnitude distribution of the events reported by the LASA Summary Bulletin, LASA does not start missing substantial numbers of earthquakes until body wave magnitude 3.7. ERL reports, on the other hand, seem to miss substantial numbers of earthquakes below body wave magnitude 4.7. NORSAR's detection capability when this study was made, was comparable to ERL. (NORSAR improved considerably after the analysis).

(6) The LASA Summary Bulletin locates earthquake epicenters within a few degrees. Distance error is twice as large as azimuth error. The NORSAR Summary Bulletin shows definite large biases in their epicenter locations. (it is expected that these biases will be removed with improved station corrections).

(7) On the basis of ERL reported events NORSAR detects a small fraction of the later phases that LASA detects.

(8) About half of the detections in either LASA or NORSAR detection logs are false alarms due to spurious noise. The LASA false alarms are confined to detections less than 1 mμ and the NORSAR false alarms extend up to an amplitude of 2 mμ.

(9) Discrimination of signals versus false alarms on the basis of only the information in the detection logs is difficult for any automatic system without the assistance of an analyst who can examine the waveforms.

Complete automatic discrimination is feasible provided one is willing to sacrifice the detection of low magnitude events.

(10) Automatic identification of later phases using a single array is definitely feasible, though the presence of an analyst to verify the identifications would improve the performance. Travel time, azimuth, distance, and amplitude information are useful in the identification of the phases--the first two being the most valuable. About 7 real later phases a day were found. On the basis of the training set, there were practically no misidentifications. The phase identifier picked two fictitious later phases a day due to detections coming in accidently in the correct sequence. The number could be halved by raising the decision threshold without losing more than one real later phase a day. The phase identifier requires the detection of the P or PKP phase in order to find the later phase.

(11) Identification of later phases with two seismic arrays is easier since it is not necessary to detect the P or PKP arrival. Performance of the two-array phase identifier was comparable to the single-array identifier and will probably improve with the implementation of a new detection processing in the NORSAR. Eight earthquakes a day were found common to LASA and NORSAR detection logs--one earthquake presumably fictitious. There were

very few misidentified later phases on the basis of the training set. Epicenter locations with the two-array phase identifier were comparable in accuracy to those of the LASA Summary Bulletin. The accuracy can be improved to almost ERL quality if additional phases to the same earthquake are detected by either array.

References

Capon, J. (1970). Analysis of Rayleigh-wave multipath propagation at LASA. Bull. Seism. Soc. Am. 60, 1701-1731.

Davies, D. and D.P. McKenzie (1969). Seismic travel-time residuals and plates. Geophys. J. Roy. Astron. Soc., 18, 51-63.

Farrel, E.J. (1971). Sensor array processing with channel recursive Bayes technique. Geophysics 36, 822-834.

Greenfield, R.J. and R.M. Sheppard (1969). The Moho depth variations under the LASA and their effects on dT/dΔ measurements. Bull. Seism. Soc. Am. 59, 409-420.

IBM data processing technique, (1959). Random Number Generation and Testing. GC-20-8011-0.

IBM final report (1972). Large-Aperture Seismic Array Signal Processing Study.

IBM final technical report (1971). Integrated Seismic Research Signal Processing System. ESD-TR-72-139.

IBM seventh quarterly technical report (1970). Integrated Seismic Research Signal Processing System. ESD-TR-72-128.

Lacoss,R.T., E.J. Kelly, and M.N. Toksoz (1969). Estimation of seismic noise structure using arrays. Geophysics 34 21-38.

Larner,K.L. (1970). Near-receiver Scattering of Teleseismic Body Waves in Layered Crust-Mantle Models having Irregular Interfaces. Ph.D. Thesis, M.I.T.

Mack, H. (1969). Nature of short period P-wave signal variations at LASA. J. Geophys. Res., 74 3161-3170.

Middleton,D. (1960). Introduction to Statistical Communication Theory, McGraw-Hill.

Richter, C.F. (1958). Elementary Seismology, W.H. Freeman & Co.

Sebestyen, G.S. (1962). Decision-making Processes in Pattern Recognition. MacMillan.

Shlien, S., and M.N. Toksoz (1970a). Frequency magnitude statistic of earthquake occurrences. _Earthquake Notes_ _41_, 5-18.

Shlien,S. and M.N. Toksoz (1970b). A clustering model for earthquake occurrences. _Bull_. _Seism_. _Soc_. _Am_. _60_,1765-1787.

Van Trees,H. (1968). _Detection Estimation and Modulation Theory_. John Wiley & Sons.

Appendix A

Criterion for Matching Predicted Signals to the Detection Log

In the evaluation of the capabilities of the arrays (Chapter 2) and in the generation of a training set, it was necessary to match predicted or reported phases to the detections in the detection log. This section describes the matching criterion that was used and the errors that were involved.

Signals were matched to detections on the basis of their arrival times. If the predicted arrival time of a phase coincides exactly with the time of the reported detection in the log, then they are perfectly matched. Usually there is a time difference between the predicted and the observed arrival times. The errors are due to several reasons. The predicted arrival time of a phase can be off by many seconds. To predict the arrival time of a signal exactly, we must know the epicenter coordinates and depth and the travel time distance depth relation exactly. Due to the lateral inhomogeneities in the earth, neither the epicenter nor travel times can be determined precisely. Secondly the Detection Processor will not always trigger at the true arrival time of the signal. If the signal is emergent, the beginning of the signal will be missed. If the signal is very strong, it will trigger the misdirected beams before the true beam. In other words the Detection Processor will trigger before the signal had propagated across the whole array. For these reasons the matching criterion involved used a finite time window.

The time window should be neither too large nor too small. If it is too large then the probability of making a bad match (e.g. signal matched to noise) is substantial. If it is too small there is a sizeable chance of missing the signal. The matching criterion used in this study generally accepted anything in the time interval of plus or minus 40 seconds of the predicted arrival time. This window was found to be more than adequately large. When the newer data is analyzed the window will probably be shortened to 15 seconds.

The probability of making a bad match may be estimated assuming a Poisson model. There are about 300 detection groups a day in the LASA detection log and 70 detection groups a day in the NORSAR log. Since weak signals are mostly false alarms all the LASA detections with MSTA < 100 and NORSAR detections with MSTA < 300 were ignored. This reduces the detection rates to 80 groups a day for LASA and 65 groups a day for NORSAR. If 100 detection groups occur on the average in one day, then the mean recurrence time is 864 seconds. In any random interval of 100 seconds in the detection log the probability of finding no detection groups is exp(-100/864) = 0.89 . This implies that the probability of a false match is less than 11%. This effect may make LASA and NORSAR to appear to have slightly better detection capability than actually.

Appendix B

Distance and Azimuth Resolution of a Large-Aperture Seismic Array

B.1 Introduction

It is important to know the resolution capability of a seismic array in the construction of an automatic phase identifier. On account of the limited aperture of a seismic array, an array can very rarely locate an earthquake to less than 1 degree error. The size of the error is very strongly dependent on the phase used to locate the event and the distance of the event. This section shows and explains how the distance and azimuth resolution of an array is related to the beam's resolution in the inverse velocity space.

B.2 Distance Resolution

In Figure B-1 the distance and azimuth of earthquakes triggering specific beams in the high resolution beam partition is plotted. Though the beams have identical resolution in inverse velocity space, it is very clear that at greater distances the region of epicenters that can trigger the same beam becomes much more spread out. This is due to the nonlinear transformation between distance and travel time.

For purposes of argument we shall stick to the standard seismic notation. Let T be the travel time of a phase from an earthquake at distance Δ . A seismic array basically observes the inverse phase velocity $\frac{dT}{d\Delta}$ by measuring the time for the seismic signal to cross the array. LASA for instance can measure $\frac{dT}{d\Delta}$ to a resolution of .15 seconds per degree using its fine beam partition. The $\frac{dT}{d\Delta}$ is directly related to the angle of incidence of the seismic signal at the array. For example if the signal is coming vertically then the signal will be observed at all seismographs simultaneously. The $\frac{dT}{d\Delta}$ depends on the distance of the event and the type of phase.

For most seismic phases there is a one to one correspondence between $\frac{dT}{d\Delta}$ and Δ the distance of the event, $\Delta = \Delta(dT/d\Delta)$. Hence once $\frac{dT}{d\Delta}$ and the phase type is known then one has a good estimate of the earthquake distance. How well one can estimate distance depends on how sensitive $\frac{dT}{d\Delta}$ is to distance. For local earthquakes $\frac{dT}{d\Delta}$ varies very rapidly with distance. In the shadow zone $\frac{dT}{d\Delta}$ is virtually constant. If $\Delta = \Delta(dT/d\Delta)$ then the error in Δ, corresponding to an error $\delta\rho$ in $\frac{dT}{d\Delta}$ is

$$\begin{aligned} \delta\Delta &= \frac{d\Delta(dT/d\Delta)}{d(dT/d\Delta)}\,\delta\rho \\ &= \delta\rho\left(\frac{d^2T}{d\Delta^2}\right)^{-1} \end{aligned} \qquad \text{(B-1)}$$

B.3 Azimuth Resolution

The azimuth resolving power of an array depends on the $\frac{dT}{d\Delta}$ of the signal. If the signal is coming nearly vertically it is very difficult to estimate the azimuth of the signal. This is practically the situation for the phases P'P' (df branch) and PKKP (bc branch). Unfortunately $\frac{dT}{d\Delta}$ is generally small for phases at large distances so the earthquake location error becomes even more appreciable with distance.

In order to estimate analytically the error in azimuth it was assumed that the error in the inverse velocity $\bar{U}$ determination is normally distributed with zero mean and standard deviation σ. Provided that there are sufficient beams covering the signal region, then this is a reasonable assumption. Let u be the magnitude and α the azimuth of the actual inverse velocity $\bar{U}$. Let ν be the measured inverse velocity and β the measured azimuth of $\bar{U}$. Then by assumption the probability of measuring ν and β given u and α is

$$p(\nu,\beta|u,\alpha) = \frac{\nu}{2\pi\sigma^2} \exp[-(\nu^2+u^2-2\nu u \cos(\alpha-\beta))/2\sigma^2]$$

$$0 \le \nu \le \infty \qquad \text{(B-2)}$$

$$0 \le \alpha \le 2\pi$$

(This is the same model for the radar problem of narrow band signal with additive normal noise.)

Then

$$p(\beta|u,\alpha) = \int p(\nu,\beta|u,\alpha)d\nu \qquad \text{(B-3)}$$

$$= \frac{1}{2\pi}\exp(-a_0^2)\ [1 + 2\sqrt{\pi}a_0\cos\gamma\ \frac{1+\mathrm{erf}(a_0\cos\gamma)}{2}\ \exp(a_0^2\cos^2\gamma)]$$

where $a_0^2 = u^2/2\sigma^2$ and $\gamma = \beta - \alpha$

Middleton (1960). The probability density function is bell shaped and becomes more peaked with larger a_0. Using the same notation as in section B.2 a_0 could be related to the inverse phase velocity and resolution by

$$a_0^2 = (dT/d\Delta)^2\ /2\sigma^2 \qquad \text{(B-4)}$$

The standard deviation of β was calculated numerically as a function of the parameter a_0 and is plotted in Figure

B-2. For actual signals a_0 varies from 50 to 10 for P & PKP phases using LASA's high resolution beam partition and from 20 to 4 using LASA's low resolution beam partition. The standard error of the azimuth determinations are in fair agreement with these values. This analysis neglects the effect of bad station corrections.

Appendix C

Improved Discrimination of Signals from False Alarms

In section 4.2 the problem of distinguishing false alarms from signals and estimating the number of signals detected by the arrays was discussed. It was concluded that as long as MSTA was above a certain threshold then one can preclude the detection being a false alarm. Below this threshold one could never be sure. This section shows mathematically what one could do to decide between signal and false alarm when the signal is below the threshold.

As was mentioned in 4.2, inclusion of seismicity information as a function of space and time could enhance the decision algorithm. With just MSTA information the posteriori probability of a detection being a false alarm would be written as

$$p(\mathrm{FA}|\mathrm{MSTA}) = \frac{p(\mathrm{MSTA}|\mathrm{FA})\,p(\mathrm{FA})}{p(\mathrm{MSTA})} \qquad \text{(C-1)}$$

where FA stands for false alarm. All the quantities on the right hand side could be estimated from the seismic and aseismic beams using the detection log as described in section 4.2. Cumulative distributions for LASA MSTA are shown in Figure E-1. p(FA) was estimated to be 0.5 for all LASA detections.

If one includes beam information then the test could be easily refined one step further. It is safe to assume that p(MSTA|FA) and p(MSTA) are independent of the beam. On the contrary p(FA) depends on the beam. If the beam is pointed to an aseismic area there would be very few signals. Thus the test could be rewritten as

$$p(FA|\,MSTA,NBM) = \frac{p(MSTA|\,FA)\,p(FA|\,NBM)}{p(MSTA)\;p(NBM)} \qquad (C-2)$$

p(FA|NBM) and p(NBM) could be estimated from the detection log. (A biased estimate p(FA|NBM) could be made by counting the number of detections above and below a certain detection threshold for a given beam. To remove the bias one must be able to estimate the percentage of signals below the MSTA threshold which depends on the beam number.)

The final step is to include time information. False alarms come at completely random times and random beams. Earthquakes are not completely random. A large earthquake generates many aftershocks. If more than two beam detections are observed within an interval of several hours, it would be less likely that they are false alarms. For one specific beam the mean recurrence time of false alarms is a little more than a day. To incorporate this time information the discriminator would count the number of detections in that beam within a time interval of t hours. Assuming that false alarms can be approximated by a Poisson model the probability of n detections occurring within a period of t hours is determined by

$$p(n) = \frac{k^n}{n!} \exp(-kt) \tag{C-3}$$

where k is the mean rate of false alarms. If p(n) is very small the discriminator will lower the MSTA threshold to accept only the normal number of false alarms.

Appendix D

Shuffling a Detection Log

In this section we describe how the synthetic detection log was generated. The synthetic detection log had to have all the statistical properties of the original detection log except that detection groups must be completely independent of each other. To obtain such a log it was decided to shuffle the beam numbers and MSTA values of the original log. Care was taken to preserve the detection groups. A group of say ten detections triggered by one signal was moved all together. For convenience a group was defined by the rule as any detection coming within 20 seconds of the previous detection belongs to the same group. Detections in a group in the new shuffled log were always spaced one second apart to avoid problems of any groups overlapping resulting in a loss of chronological order. These simplifications would still give a good approximation to a random detection log.

The detection beams and MSTA values were always shuffled by the same algorithm. A set of 306 or less detection groups in chronological order was read into core. The order of the groups was randomized by the following algorithm.

$$j = 5^{i} \quad \text{(modulo 307)} \qquad \text{(D-1)}$$

where i was the original position and j is the new position. If there were only m detection groups in the original set where m is less than 306, then all j's greater than m were ignored and i was incremented without discarding that detection. m was usually less than 306 since once the total number of detections was equal to 306 no new detection groups were read in. Because 5 is a primitive root of order 306 in modulo 307 the j elements would never repeat as i went from 1 to 306. The shuffled groups were then written back onto another tape.

The theory of this method is described in the IBM Data Processing Technique (1959).

Appendix E

Numerical Evaluation of l_i for the Single Array Phase Identifier

The log likelihood statistic l_i for the single array phase identifier is computed from the parameters S_1, S_2, S_3, MSTA1, MSTA2 and r. The actual formulae used in the program are given here.

Means and variance of S_1, S_2, and S_3 were determined on the basis of the distributions in Figure 21. The normal approximations to these distributions are listed in Table E-1.

Estimation of $p(\text{MSTA1, MSTA2} | H_i)$ $i = 1,2 \ldots 10$ was estimated indirectly through the intermediate parameter

$$r = \ln (\text{MSTA2/MSTA1}) \qquad \text{(E-1)}$$

(plotted in Figure 22 for the training set.) The parameter r was found to be reasonably approximated by a normal distribution

$$p(r|H_i) = \frac{1}{0.95\sqrt{2\pi}} \exp\left(-(r-\bar{r}_{H_i})^2\right)/1.80 \qquad \text{(E-2)}$$

where r_{H_i} depends on the hypothesis. The r_{H_i} values are also listed in Table E-1. Transforming parameter r to the MSTA's,

$$\begin{aligned} p(\text{MSTA2}|\text{MSTA1},H_i) &= p(r|H_i) \left|\frac{dr}{d\ \text{MSTA2}}\right| \\ &= \frac{p(r|H_i)}{\text{MSTA2}} \end{aligned} \qquad \text{(E-3)}$$

(It was implicitly assumed that the distribution of r is independent of MSTA1. This is reasonable since one may assume that percentage of attenuation suffered by a seismic signal is independent of the signal strength.) Hence $p(\text{MSTA1},\text{MSTA2}|H_i)$ was estimated using

$$p(\text{MSTA1},\text{MSTA2}|H_i) = p(\text{MSTA2}|\text{MSTA1}, H_i)p(\text{MSTA1})$$

p(MSTA1) was assumed to be the signal distribution determined from the aseismic beams. The approximation used is

$$\begin{aligned} p(\text{MSTA1}) &= p(\text{MSTA}) - p(\text{FA})p(\text{MSTA}|\text{FA}) \\ &= 270\ \text{MSTA}^{-2.3} - \frac{1}{42}\ \text{EXP}(-(\text{MSTA}-30)/42) \end{aligned} \qquad \text{(E-4)}$$

which was inferred from Figure E-1 (FA means false alarm).

For the complement hypothesis H_{11} the MSTA's of the two detections are most likely independent of each other; hence

$$\begin{aligned} p(\mathrm{MSTA1},\mathrm{MSTA2}\,|\,H_{11}) &= p(\mathrm{MSTA1})\,p(\mathrm{MSTA2}) \\ &= (135)^{2}\ (\mathrm{MSTA1})^{-2.3}\ (\mathrm{MSTA2})^{-2.3} \end{aligned} \tag{E-5}$$

where p(MSTA)is the distribution on the left of Figure E-1. In all cases if either MSTA was below 30 the detection pair was automatically rejected.

Appendix F

Spherical Surface Transformation

The transformation to convert the distance and azimuth of a beam from array 1 (D_1, A_1) to the distance and azimuth from array 2 (d_2, a_2) is given here. Fundamentally this transformation involves the solution of a spherical triangle given two sides and an included angle. The distance and the azimuth of one array is known with respect to the other. (LASA is about 60 degrees from NORSAR.) Letting Δ be the distance of array 1 to array 2, c_2 be the azimuth of array 2 with respect to array 1 and C_1 the azimuth of array 1 with respect to array 2 then

$$\cos d_2 = \cos\Delta \cos D_1 + \sin\Delta \sin D_1 \cos B_1$$

and

$$\cos b_2 = \frac{\cos D_1 - \cos\Delta \cos d_2}{\sin\Delta \sin d_2} \qquad \text{(F-1)}$$

where $B_1 = A_1 - C_1$

and $b_2 = a_2 - c_2$

Partial derivatives $\frac{\partial d_2}{\partial D_1}$, $\frac{\partial d_2}{\partial A_1}$, $\frac{\partial a_2}{\partial D_1}$, and $\frac{\partial a_2}{\partial A_1}$ needed to estimate the standard errors of the new coordinates from the old were obtained by straight differentiation. For example

$$\frac{\partial d_2}{\partial D_1} = \frac{\cos\Delta \ \sin D_1 - \sin\Delta \ \cos D_1 \ \cos B_1}{\sin d_2}$$

$$\frac{\partial d_2}{\partial A_1} = \frac{\sin\Delta \ \ \sin D_1 \ \sin B_1}{\sin d_2}$$

$$\frac{\partial a_2}{\partial D_1} = \frac{\sin D_1 - \cos\Delta \ \ \sin d_2 \ \frac{\partial d_2}{\partial D_1}}{\sin\Delta \ \ \sin d_2 \ \sin b_2} \qquad \text{(F-2)}$$

$$- \frac{\cos D_1 - \cos\Delta \ \ \cos d_2}{\sin\Delta \ \ \sin^2 d_2 \ \sin b_2} \ \cos d_2 \ \frac{\partial d_2}{\partial D_1}$$

$\frac{\partial a_2}{\partial A_1}$ same as $\frac{\partial a_2}{\partial D_1}$ except substitute $\frac{\partial d_2}{\partial A_1}$ for $\frac{\partial d_2}{\partial D_1}$

A two dimensional plot of $\frac{\partial d_2}{\partial D_1}$ is given in Figure F-1. (Minimum and maximum values of the transformation are plus and minus 1.)

Appendix G

Estimation of Λ_{ij}

The estimation of Λ_{ij} from the parameters SA, SB, and SC (identifiers L and N have been suppressed here since they are not necessary) is much more involved because the variance of these parameters depend on the phase types and the epicenter coordinates. The variance of these parameters depends on the resolution of the beam, the inverse phase velocity of the beam, the derivatives of the travel time curves of the phases, and the spherical geometry. For purposes of approximation it shall be assumed that all azimuth and distance errors of a beam are independent Gaussian variables.

The resolution of the beam in inverse velocity space was assumed to depend on only the beam partition. LASA and NORSAR each have two overlapping partitions of beams. The resolution of these different beam partitions were estimated from the training set. Letting Δp be the standard error of the beam in inverse velocity space then it follows from Appendix B that the standard errors in distance and azimuth of the beam is

$$\Delta D = \Delta p(d^2T/d\Delta^2)^{-1} \tag{G-1}$$

and $\Delta A = q(a_0)$

where $a_0^2 = (dT/d\Delta)^2/2p^2$

and q is the function plotted in Figure B-2. The derivative $\frac{dT}{d\Delta}$ depends on the presumed phase type and the distance of the event.

The parameters SA and SB were determined from both LASA and NORSAR beam detection parameters; therefore the variances of these two parameters depend on the variances of ΔD and ΔA for the two beams. The coordinates of one of the beams has been transformed to a new frame of reference. This requires the estimation of the covariance matrix of the beam coordinates in the new system. Though the ΔD and ΔA errors were independent of each other in the old system, the errors in the new coordinate system are definitely coupled. (To imagine how dramatically these errors can change consider the situation of a LASA beam pointed in the vicinity of NORSAR. What are the effects of errors in distance and azimuth of the LASA beam on the azimuth of that beam with respect to NORSAR?) Linearizing the transformation locally about the coordinates of interest it follows that provided the variances $\mathrm{var}(D_1)$ and $\mathrm{var}(A_1)$ of the old coordinates are not too large then the variances of the new coordinates are given by

$$\begin{aligned}
\mathrm{var}(d_2) &= \mathrm{var}(D_1)\,\frac{\partial d_2}{\partial D_1}\,\frac{\partial d_2}{\partial D_1} + \mathrm{var}(A_1)\,\frac{\partial d_2}{\partial A_1}\,\frac{\partial d_2}{\partial A_1}\\
\mathrm{var}(a_2) &= \mathrm{var}(D_1)\,\frac{\partial a_2}{\partial D_1}\,\frac{\partial a_2}{\partial D_1} + \mathrm{var}(A_1)\,\frac{\partial a_2}{\partial A_1}\,\frac{\partial a_2}{\partial A_1}\\
\mathrm{covar}\ (a_2,d_2) &= \mathrm{covar}\ (d_2,a_2) \qquad \text{(G-2)}\\
&= \frac{\partial a_2}{\partial D_1}\mathrm{var}(D_1) + \frac{\partial d_2}{\partial A_1}\mathrm{var}(A_1)
\end{aligned}$$

where the partials are obtained from the spherical transformation given in Appendix G.

The covariance matrix of the parameters SA and SB can now be easily evaluated using the fact that the covariance matrix of the difference of independent Gaussian vectors is the sum of the covariance matrices of the individual random vectors. Hence

$$\begin{aligned}
\mathrm{var}(SA) &= \mathrm{var}(d_2) + \mathrm{var}(D_1)\\
\mathrm{var}(SB) &= \mathrm{var}(a_2) + \mathrm{var}(A_1) \qquad \text{(G-3)}
\end{aligned}$$

and $\mathrm{covar}\ (SA,SB) = \mathrm{covar}\ (a_2,D_1)$

The variance of parameter SC depends on other factors. Recall that SC was defined to be the difference between the observed travel time interval and the expected travel time interval of the LASA and NORSAR detection pair assuming a specific phase interpretation and the epicenter being located at one of the beams. The biggest source of error of parameter SC is the uncertainty of the beam location. (Errors in measurement of detection time are negligible.) The magnitude of the error depends on how well one can estimate the distance and azimuth of the epicenter D_1, A_1 and how sensitive the travel times are to these parameters. For some geometry and phase types the errors can cancel out. As an epicenter moves away from one array it may come closer to the other array and hence the travel time interval for the distance error may be small. Estimation of the error in SC involves two major contributions, the uncertainty of distance from the first array D_1 and the uncertainty in distance from the second array d_2. The uncertainty in d_2 depends on errors in both D_1 and A_1

$$(\Delta d_2)^2 = ((\partial d_2/\partial D_1)\Delta D_1)^2 + ((\partial d_2/\partial A_1)\Delta A_1)^2 \qquad \text{(G-4)}$$

Linearizing the distance travel time relation in the area of interest, then

$$\text{var}(SC) = (\Delta D_1 \frac{dT}{dD_1} + \Delta d_2 \frac{dT}{dd_2})^2 \qquad \text{(G-5)}$$

where the travel time derivatives depend on the particular phase interpretations for the two detections. The standard error in SA and SB varied in the range of 2 to 25 degrees. The standard error of SC could be as large as 150 seconds. Approximating the distributions of SA, SB, and SC with a Gaussian model the numerator of the likelihood ratio could be easily evaluated for hypotheses H_{ij} (i,j = 1,2,...7).

Evaluation of the denominator of the likelihood ratio was much simpler. The distribution of SA, SB and SC could be estimated directly from the synthetic detection log and approximated. The distribution of SA had a larger variance than the corresponding parameters S_1 and S_2 in the single array phase identifier. The distribution of SB was comparable to S_3. SC was uniformly distributed.

Table 1

Confusion Matrix

Phase Type	Identification PcP	ScP	SKP	PP	PKKP	P'P'	Missed
PcP	52	0	0	0	0	0	0
ScP	0	15	0	0	0	0	0
SKP	0	0	10	0	0	0	0
PP	0	0	0	67	0	0	2
PKKP	0	0	0	0	36	0	3
P'P'	0	0	0	0	0	4	1

Table 2

$H_{i,j}$		Number Correctly Classified	Number Missed	Number Misclassified as
i	j			
1	1	199	5	1(5,4) 1(1,5) 1(5,1)
2	1	22	1	
3	1	11	2	
4	1	5		
5	1	48	5	1(1,1)
6	1	28	1	1(4,1)
7	1	7	4	
1	2	5		
2	2			
3	2			
4	2			
5	2	1		
6	2	1	1	
7	2		1	
1	3	1	1	
2	3			
3	3			
4	3			
5	3			
6	3			
7	3			
1	4	8		
2	4			
3	4			
4	4			
5	4	2		
6	4	1		
7	4			
1	5	15	1	
2	5	1		
3	5	1		
4	5	1		
5	5	6	1	
6	5	10	1	2(1,5) 1(1,1)
7	5		1	

Table 2
(contd.)

$H_{i,j}$		Number Correctly Classified	Number Missed	Number Misclassified as
i	j			
1	6	6	1	
2	6			
3	6			
4	6		1	
5	6	3	1	
6	6	4	1	
7	6			
1	7	5		
2	7	1		
3	7			
4	7			
5	7			
6	7	1		
7	7	2		

Code

1	P or PKP	2	PcP	3	ScP
4	SKP	5	PP	6	PKKP
7	P'P'				

Table E-1

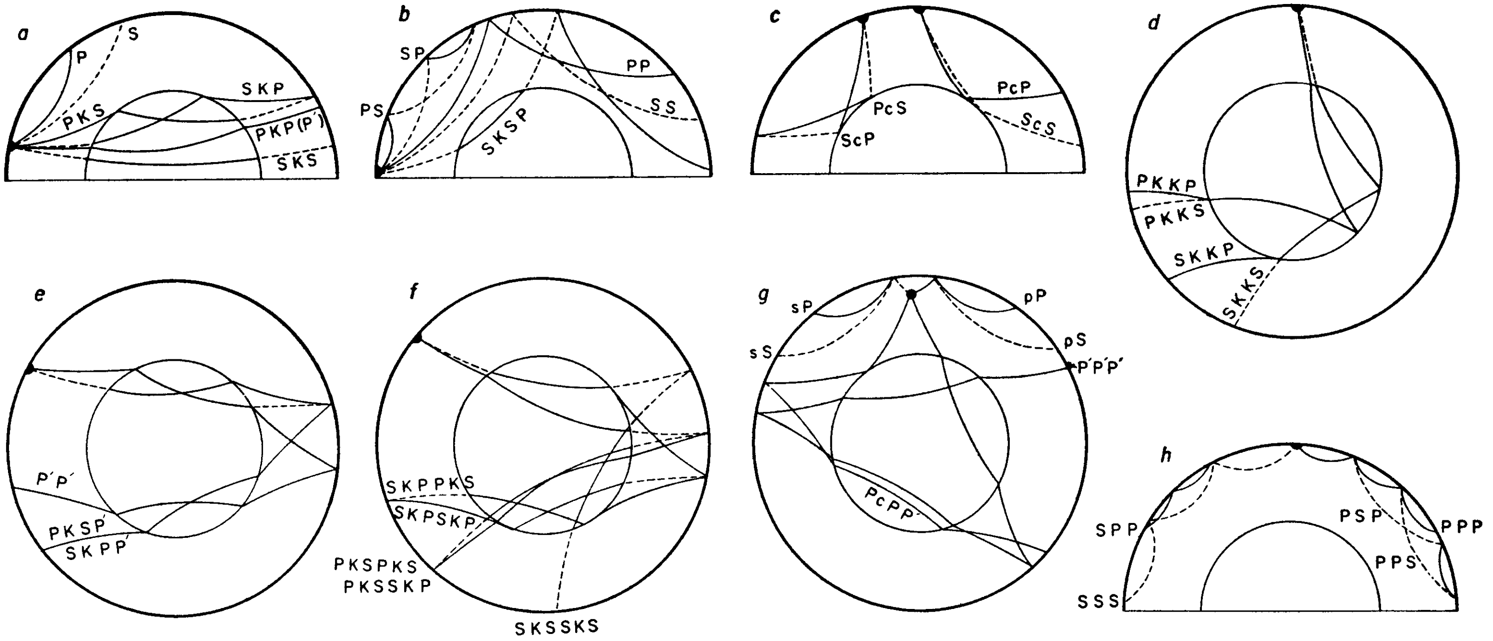
	correctly identified	unrelated detections
S_1	N(1,60)	N(-15,760)
S_2	N(1,57)	N(8,1410)
S_3	N(1,146)	uniform

$p(\mathrm{MSTA1}, \mathrm{MSTA2})$

$p(\mathrm{MSTA1},\mathrm{signal})\,p(\mathrm{MSTA2} | \mathrm{MSTA1},H_i)$ $p(\mathrm{MSTA})\,p(\mathrm{MSTA})$

N(a,b) normal, mean a and variance b

H_i	$\overline{m}_{H_i}$
P-PKP	0.0
PcP	1.1
ScP	1.9
SKP	.3
PP	.8
PKKP	.9
P'P'	2.2



Paths of body waves of teleseisms, with letter symbols. Longitudinal wave ray segments shown as full lines; transverse wave ray segments shown dashed.

Fig 1

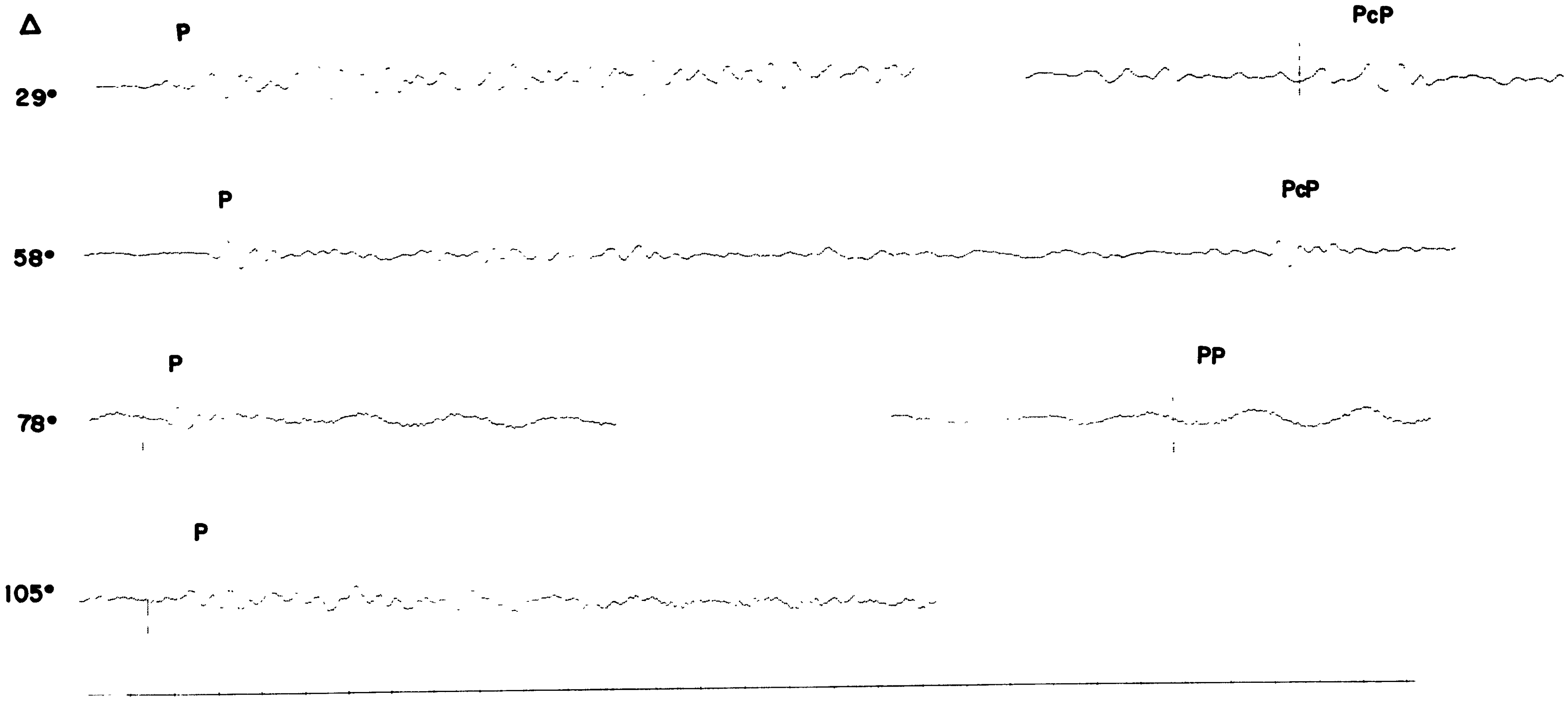
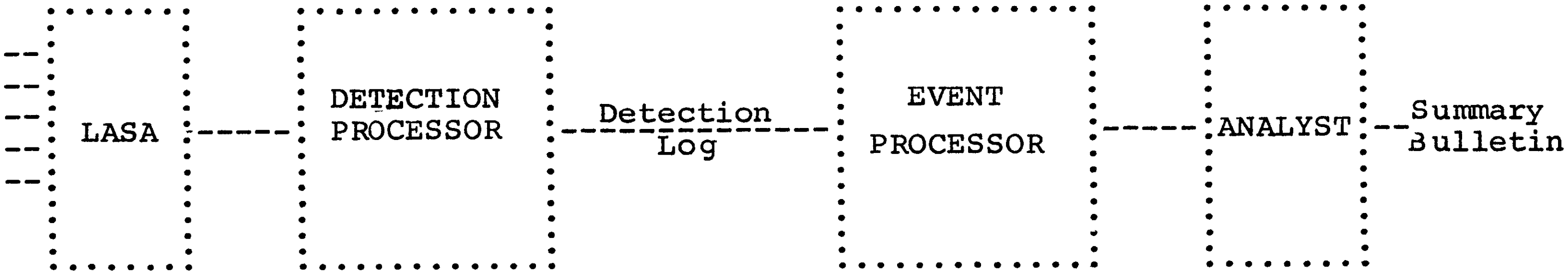


Fig 2

LASA SIGNAL PROCESSOR



DETECTION PROCESSOR

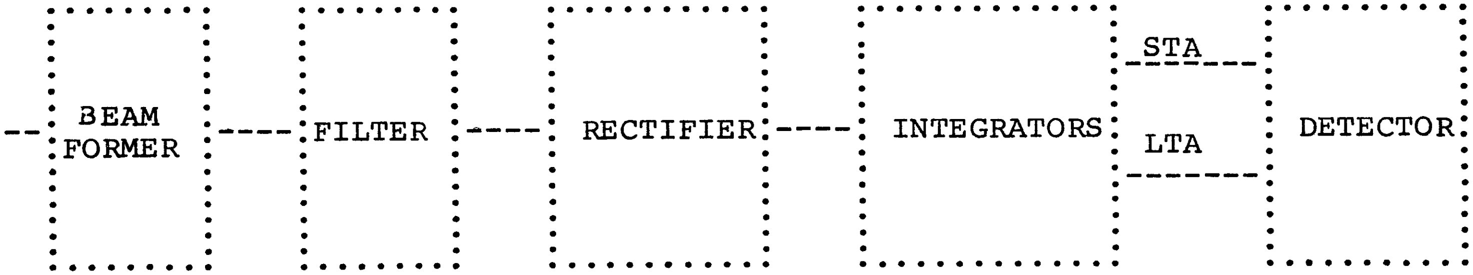
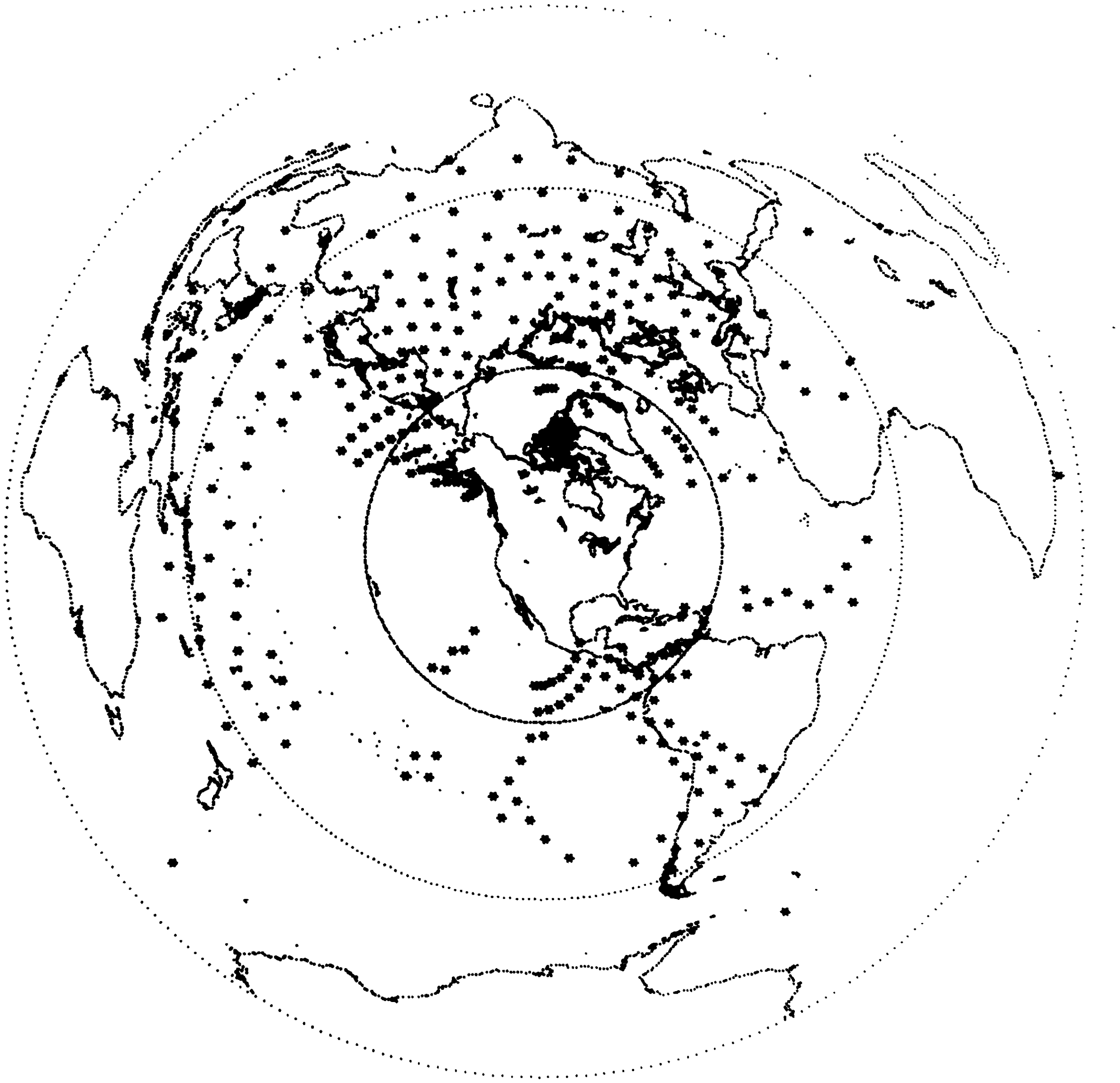


Figure 3

Fig 4



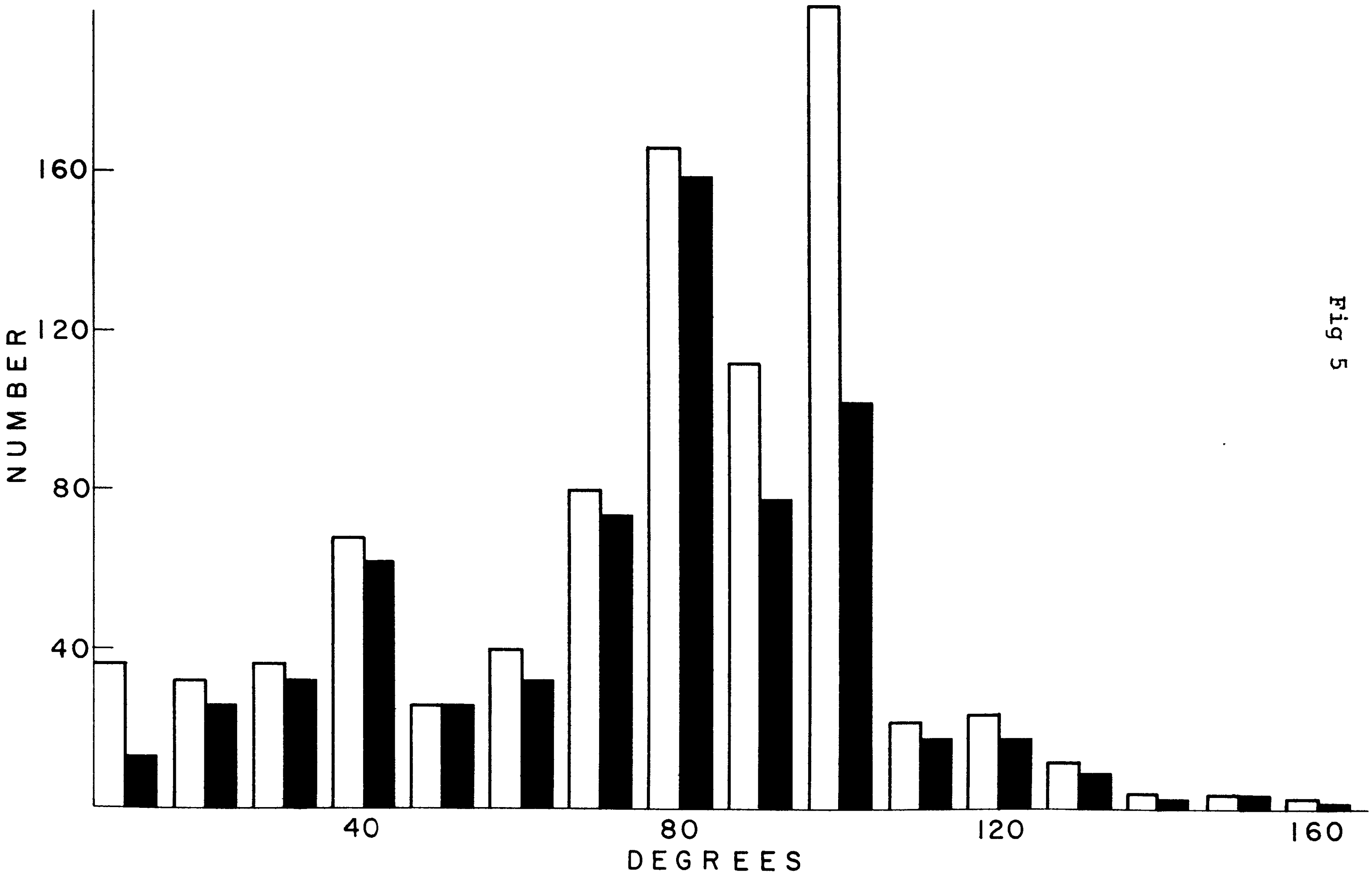


Fig 5

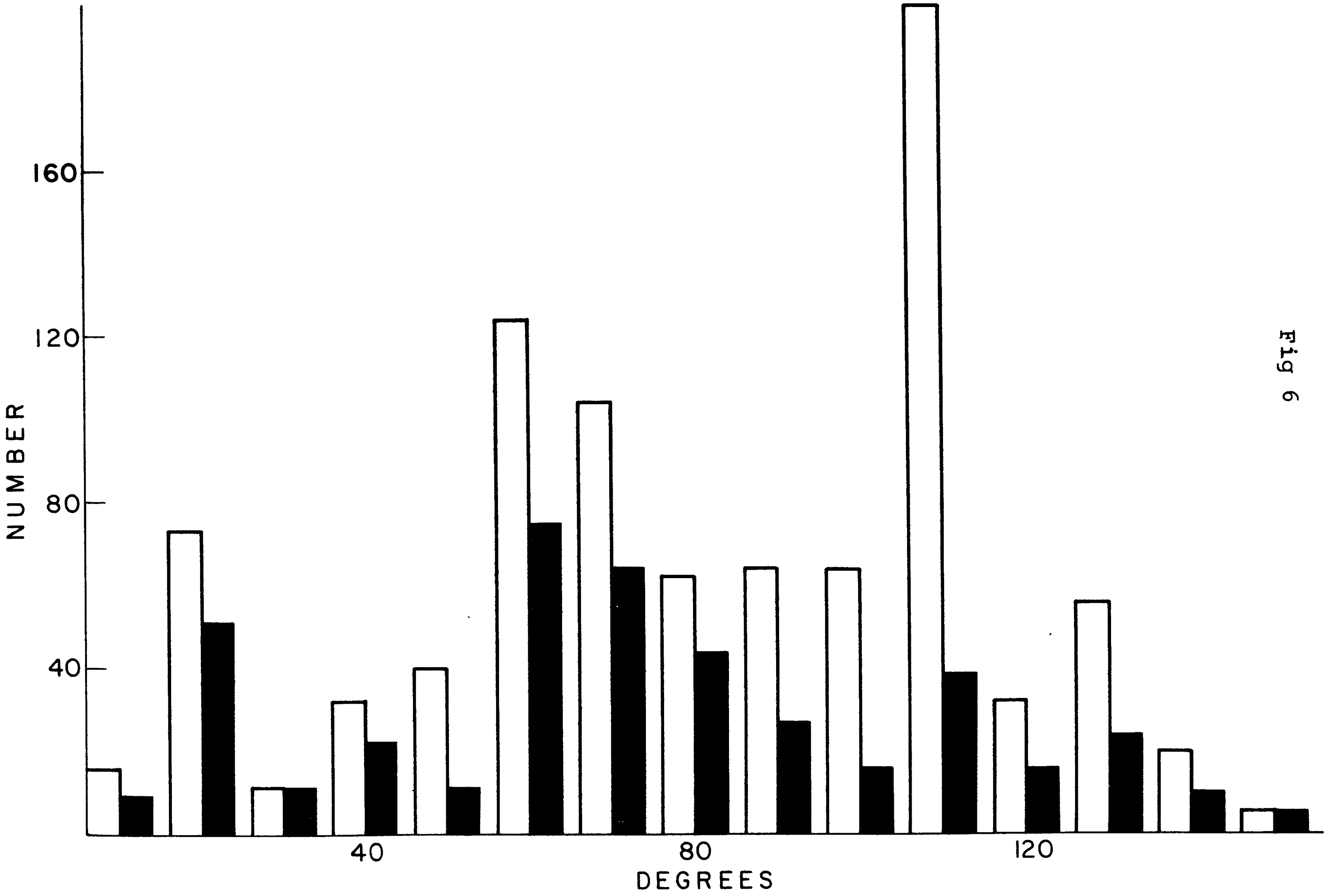


Fig 6

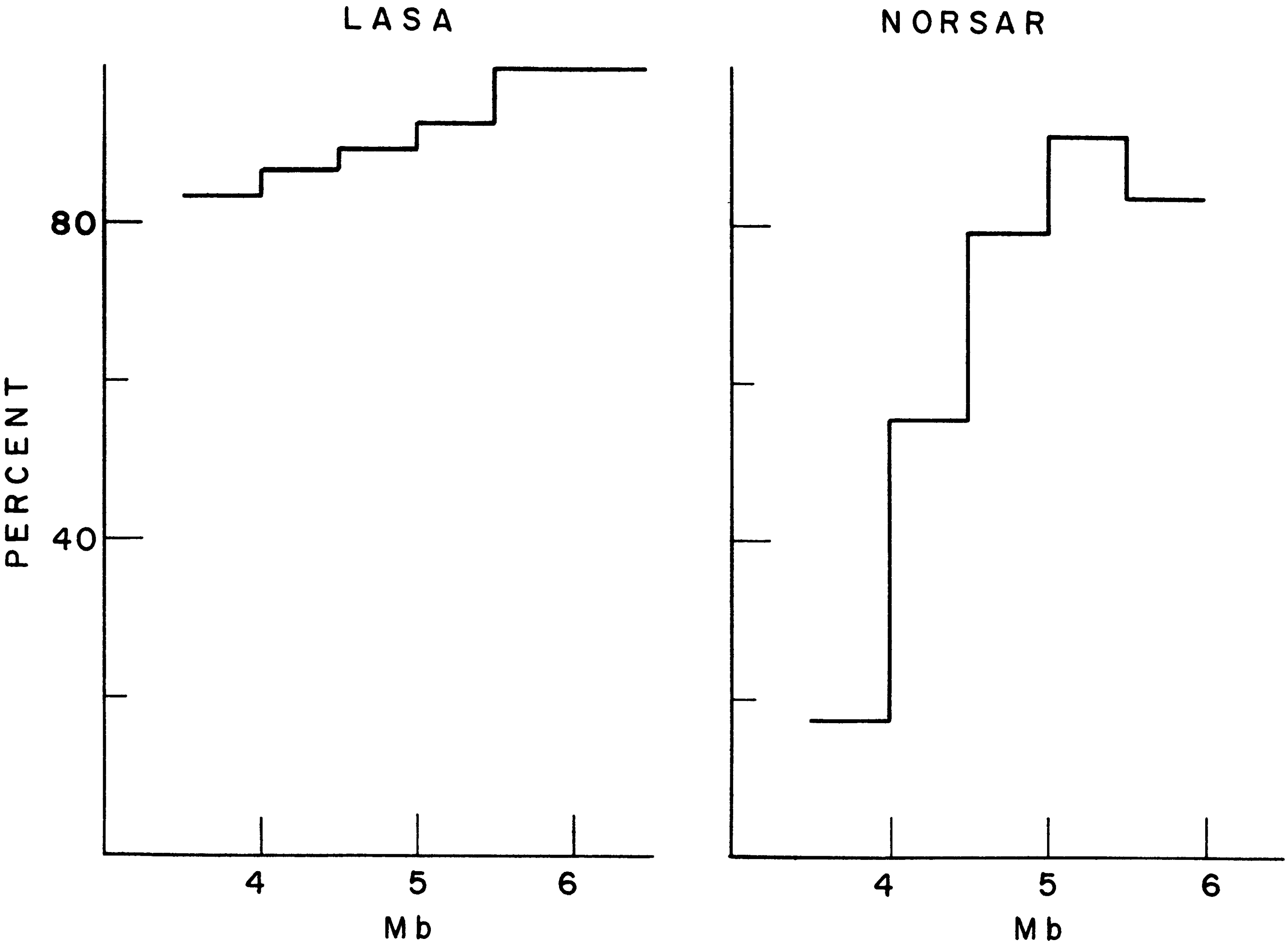


Fig 7

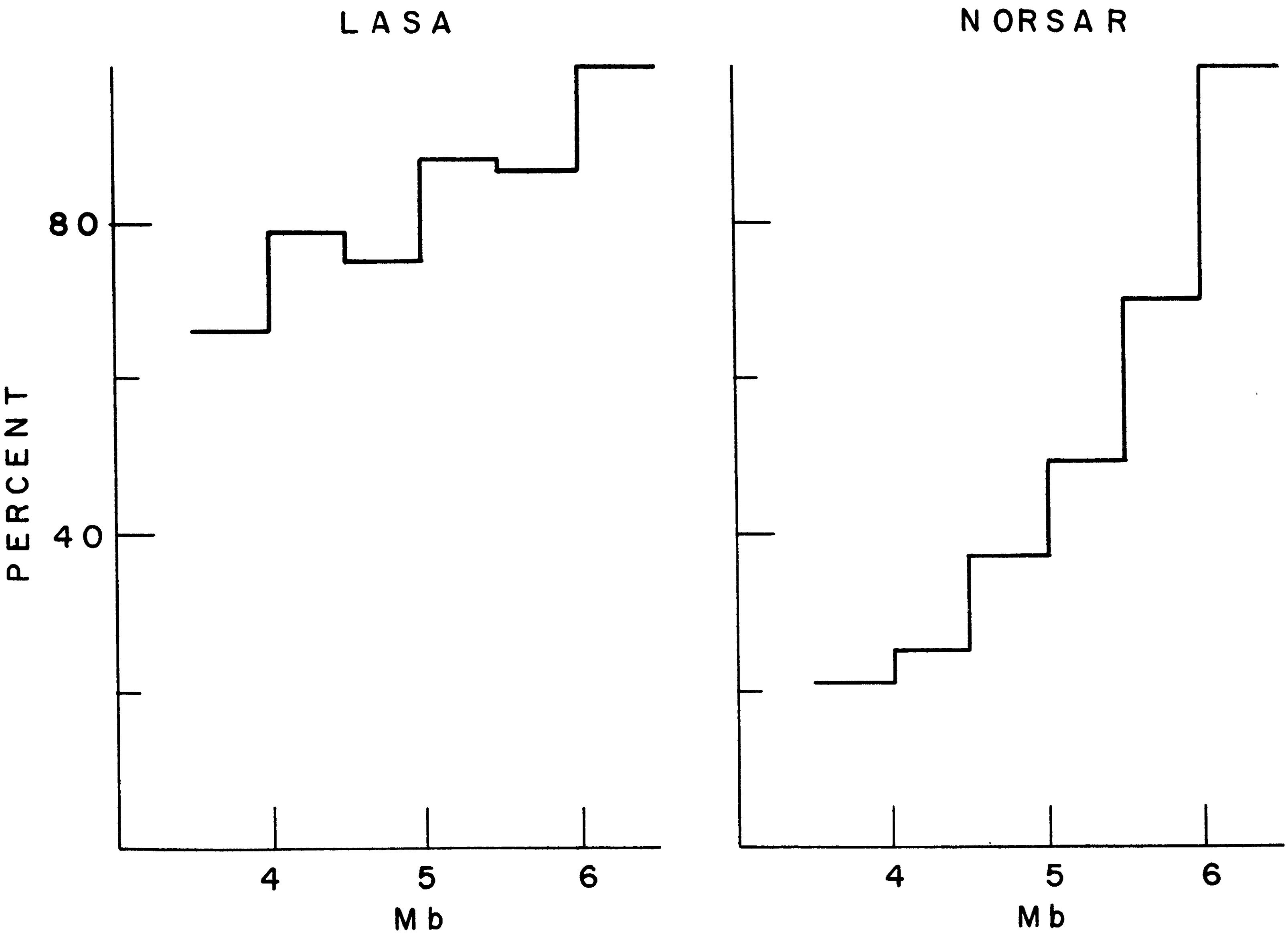


Fig 8

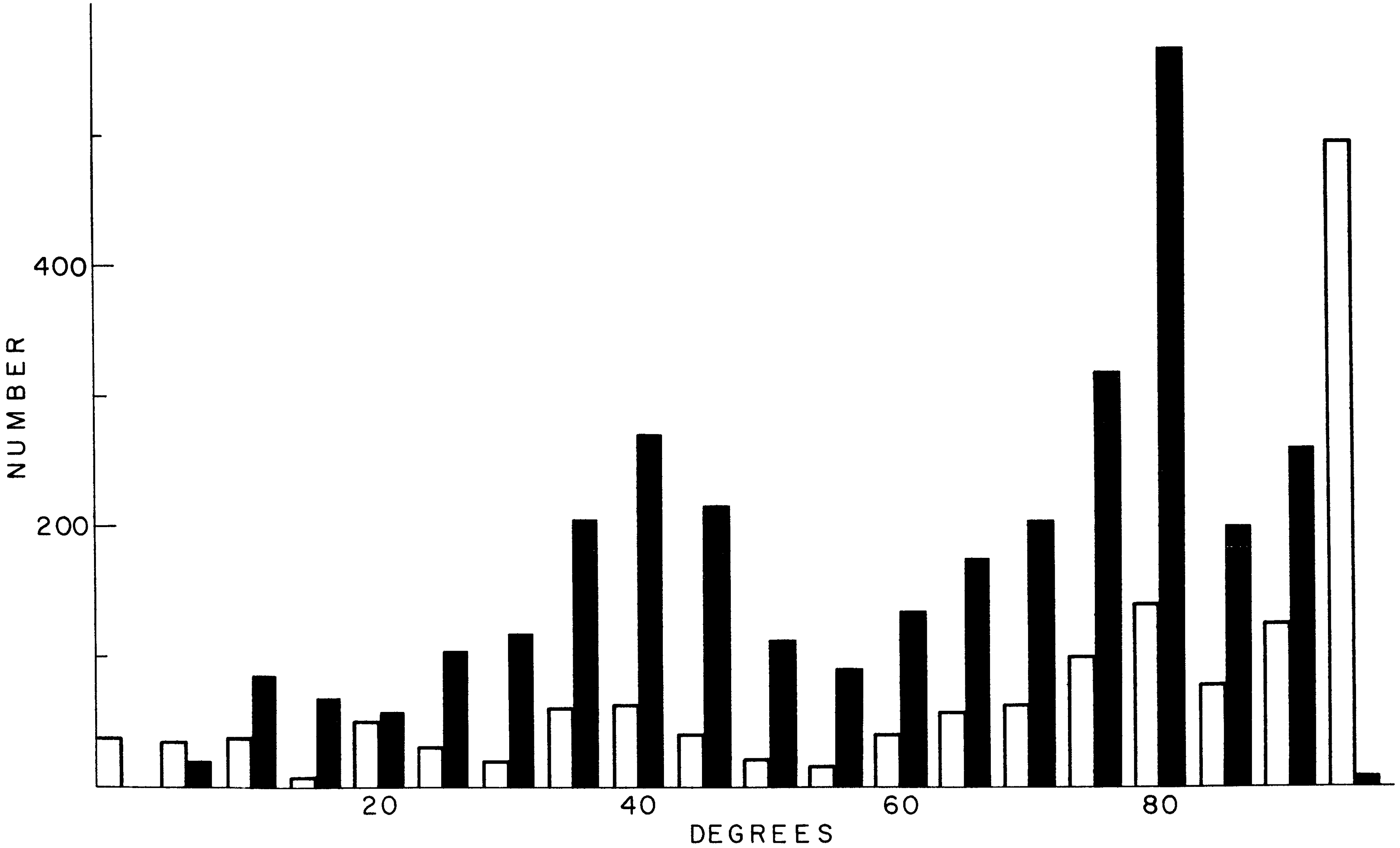
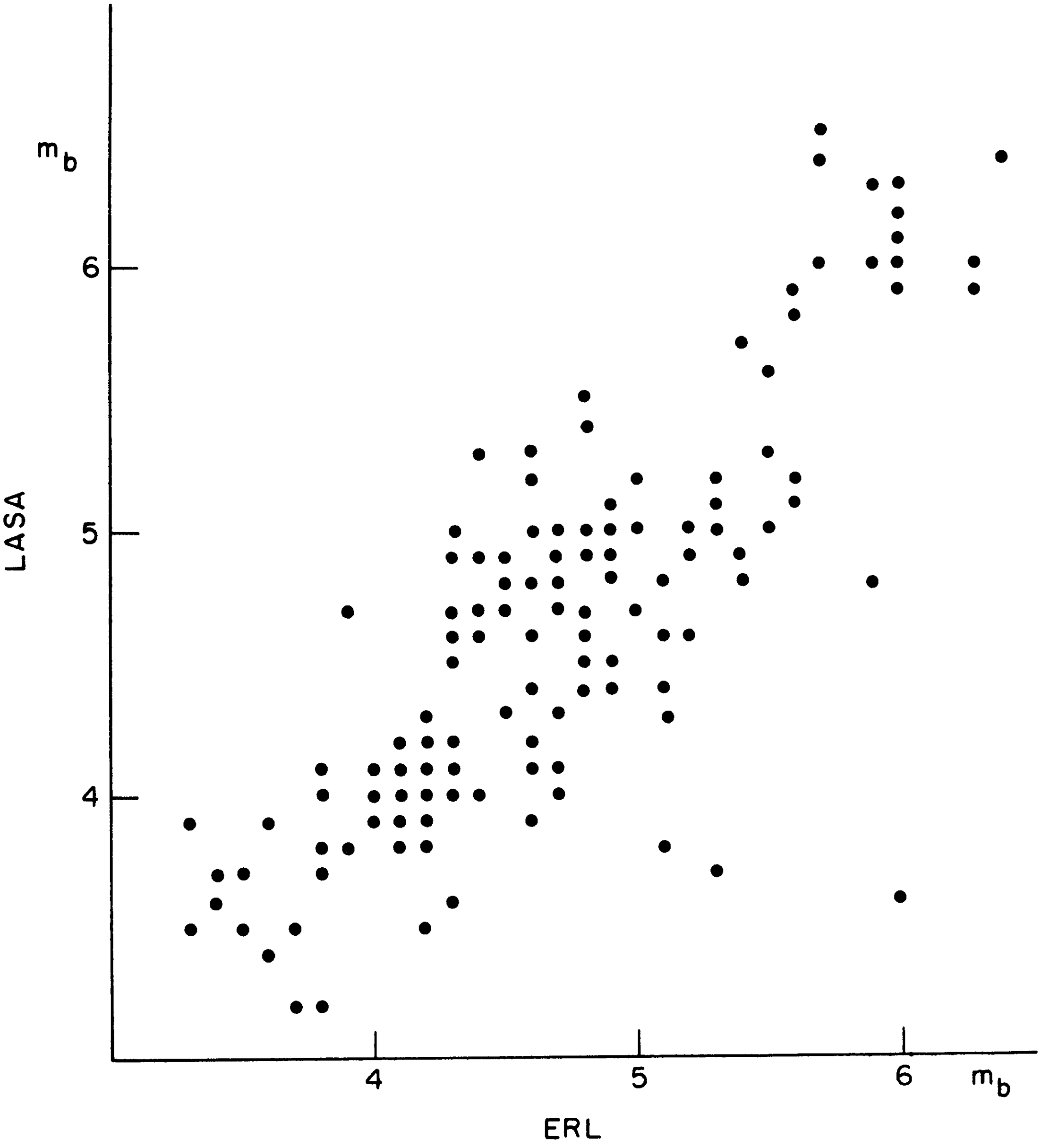


Fig 9

Fig 10



FREQUENCY - MAGNITUDE

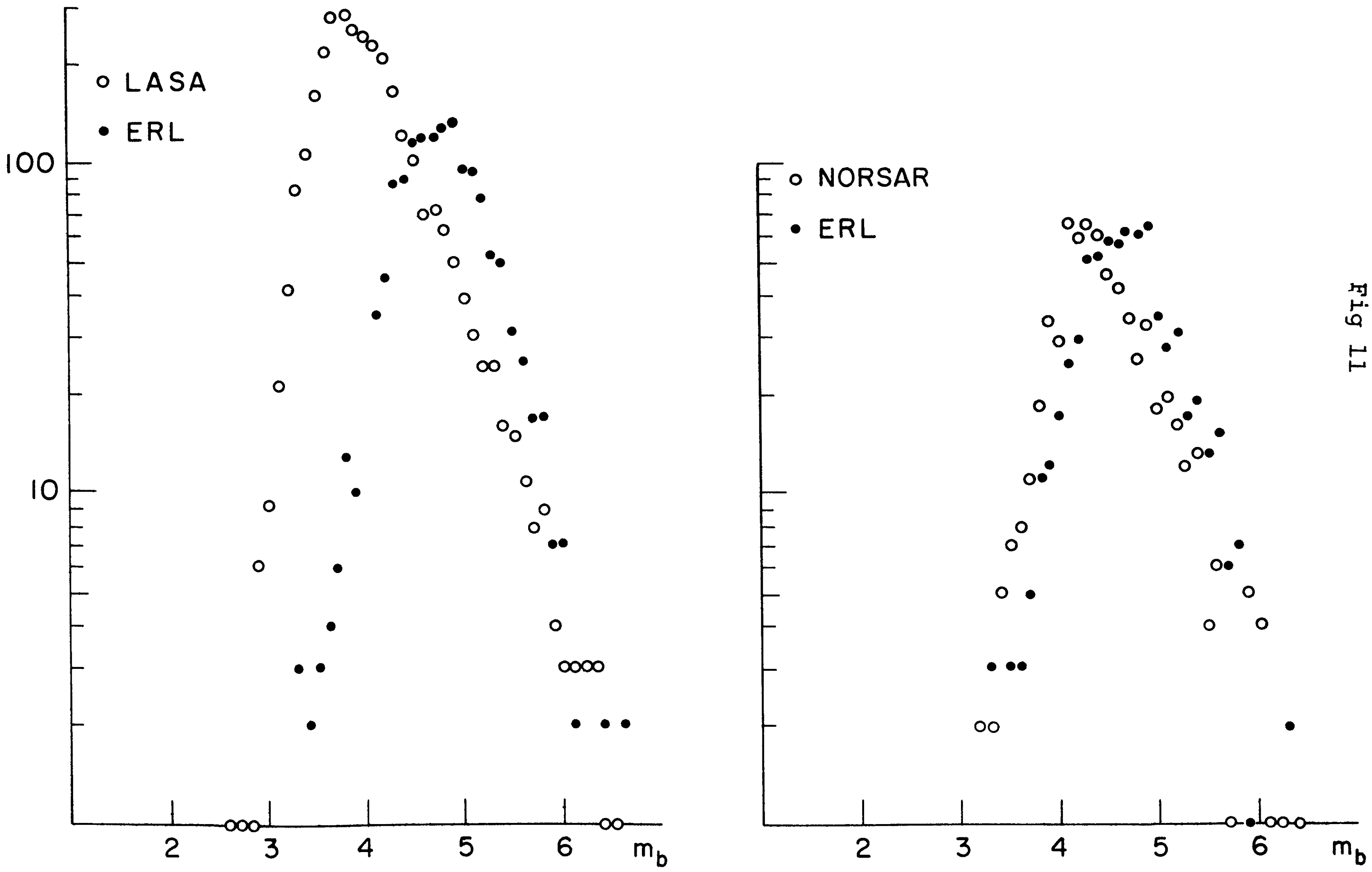


Fig 11

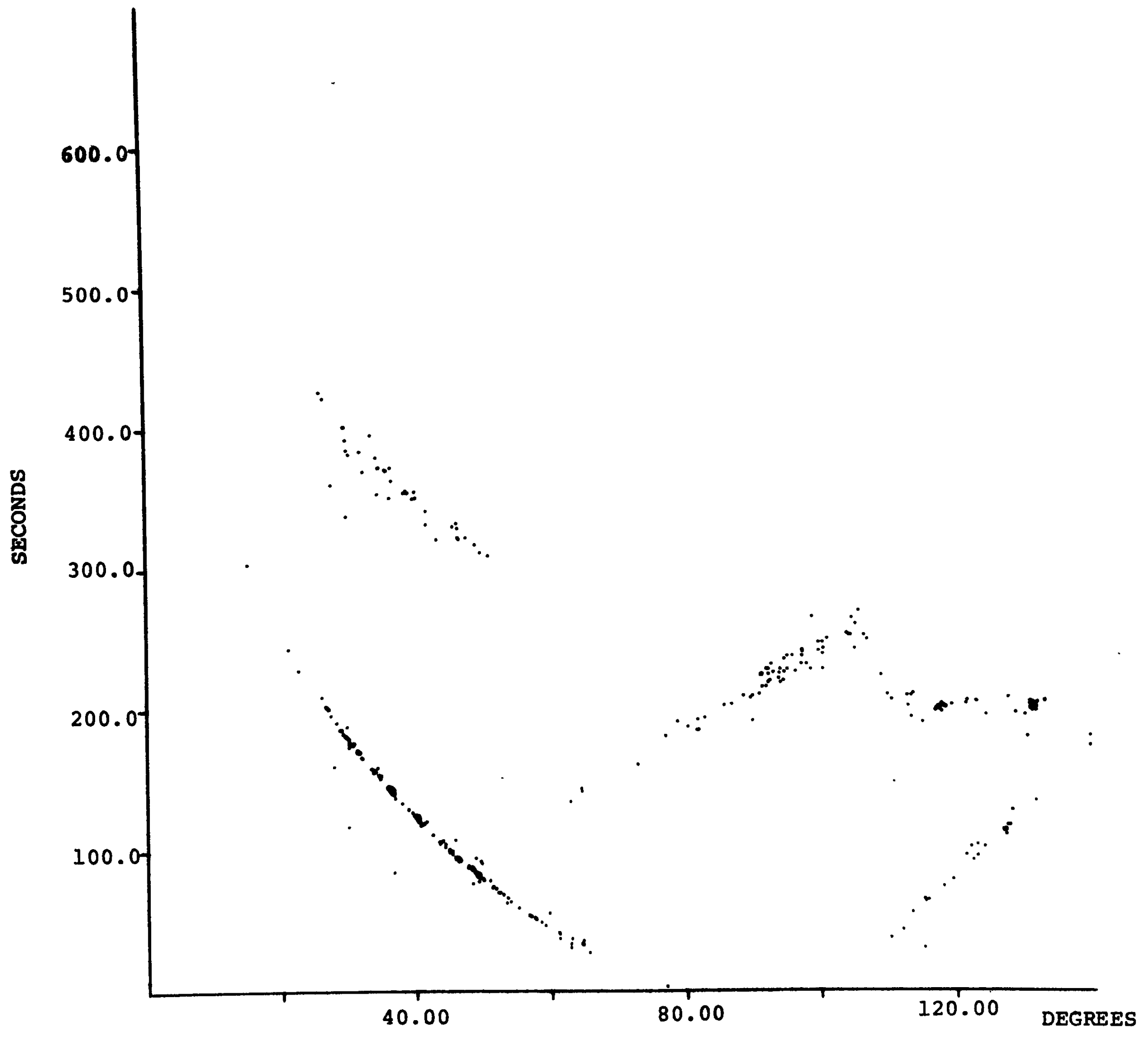

600.0
500.0
400.0
300.0
200.0
100.0
SECONDS
40.00
80.00
120.00
DEGREES

Fig 12

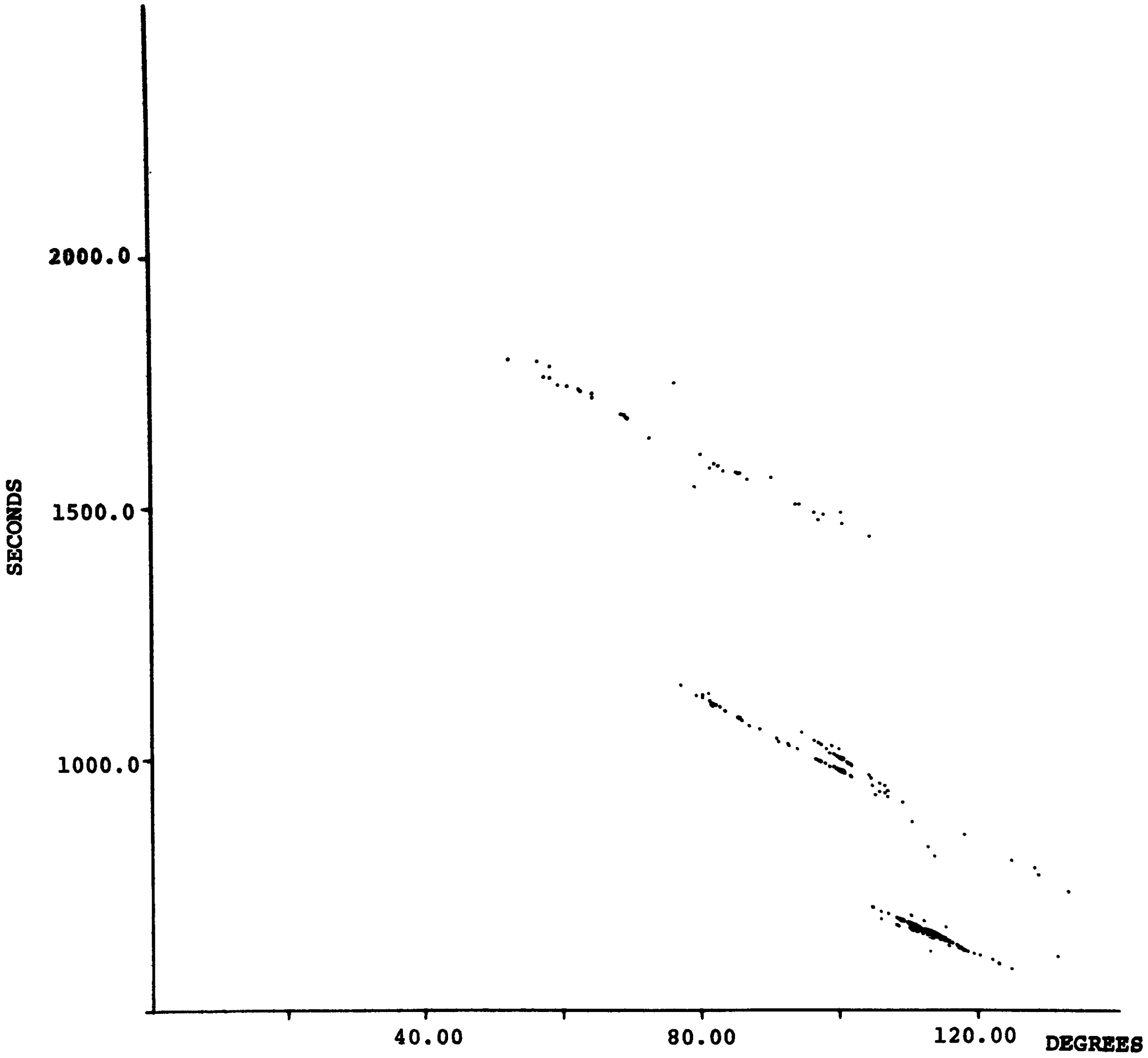


Fig 13

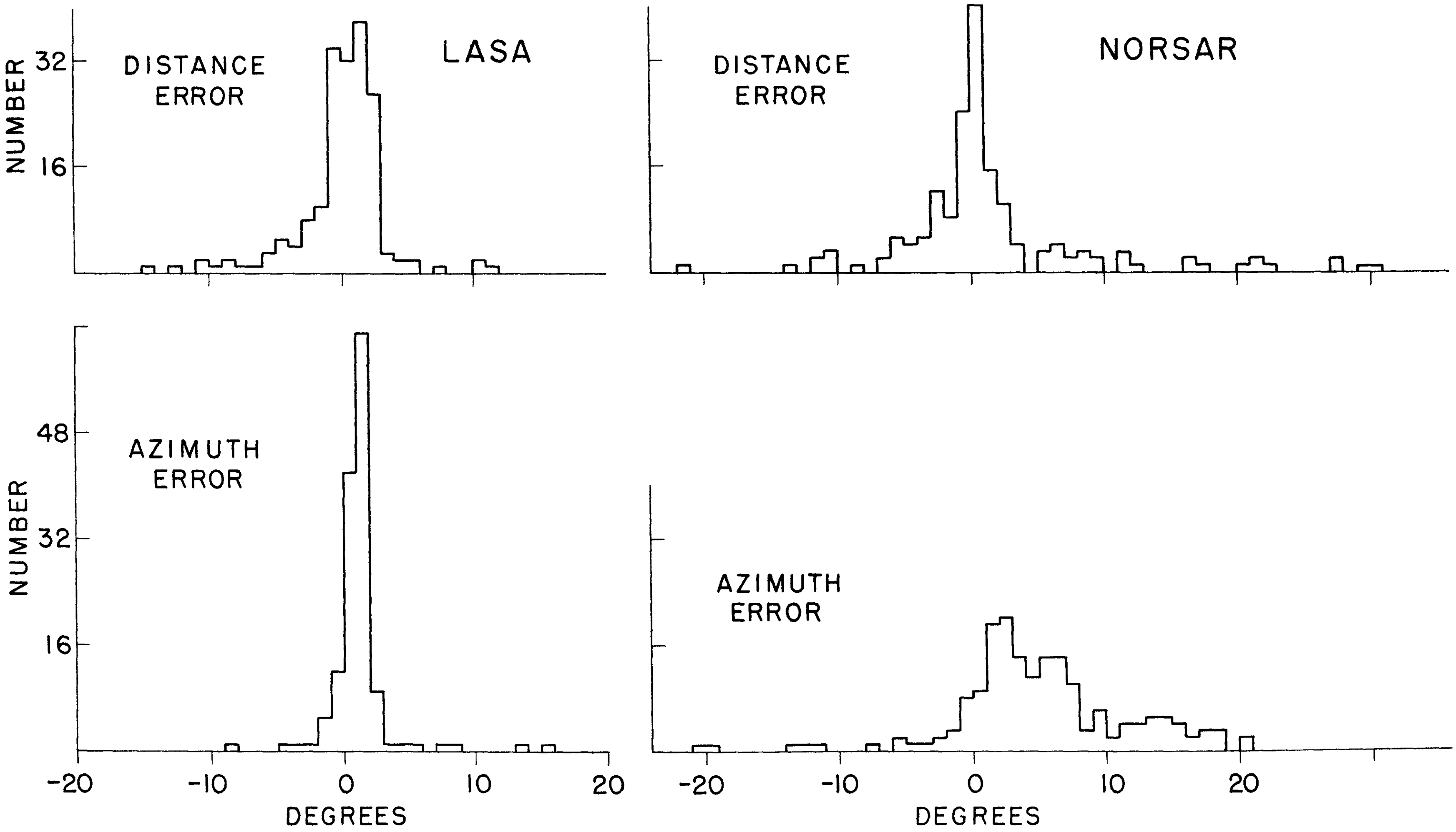
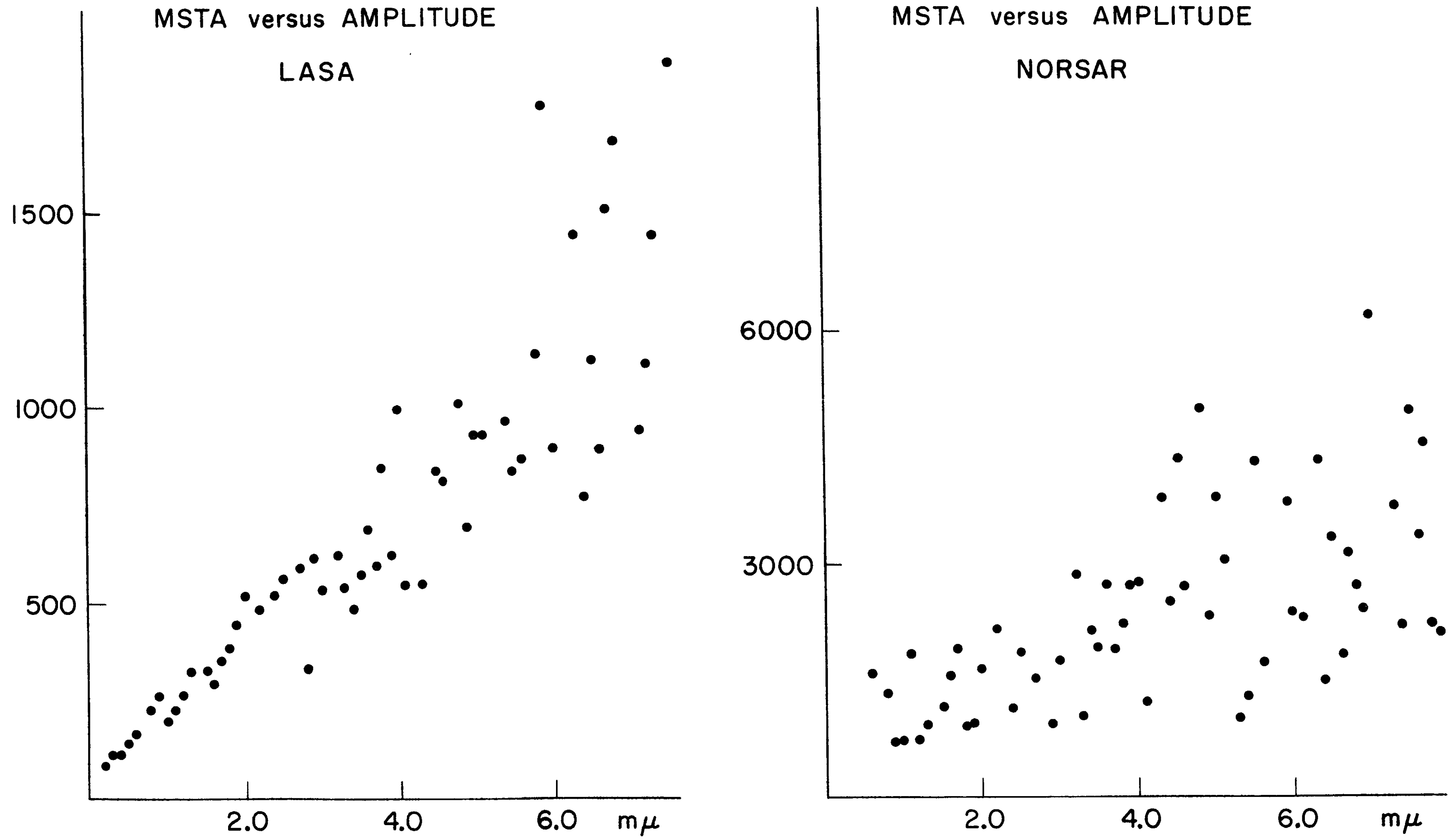

LASA
NORSAR
DISTANCE ERROR
DISTANCE ERROR
AZIMUTH ERROR
AZIMUTH ERROR
NUMBER
32
16
NUMBER
48
32
16
-20
-10
0
10
20
-20
-10
0
10
20
DEGREES
DEGREES

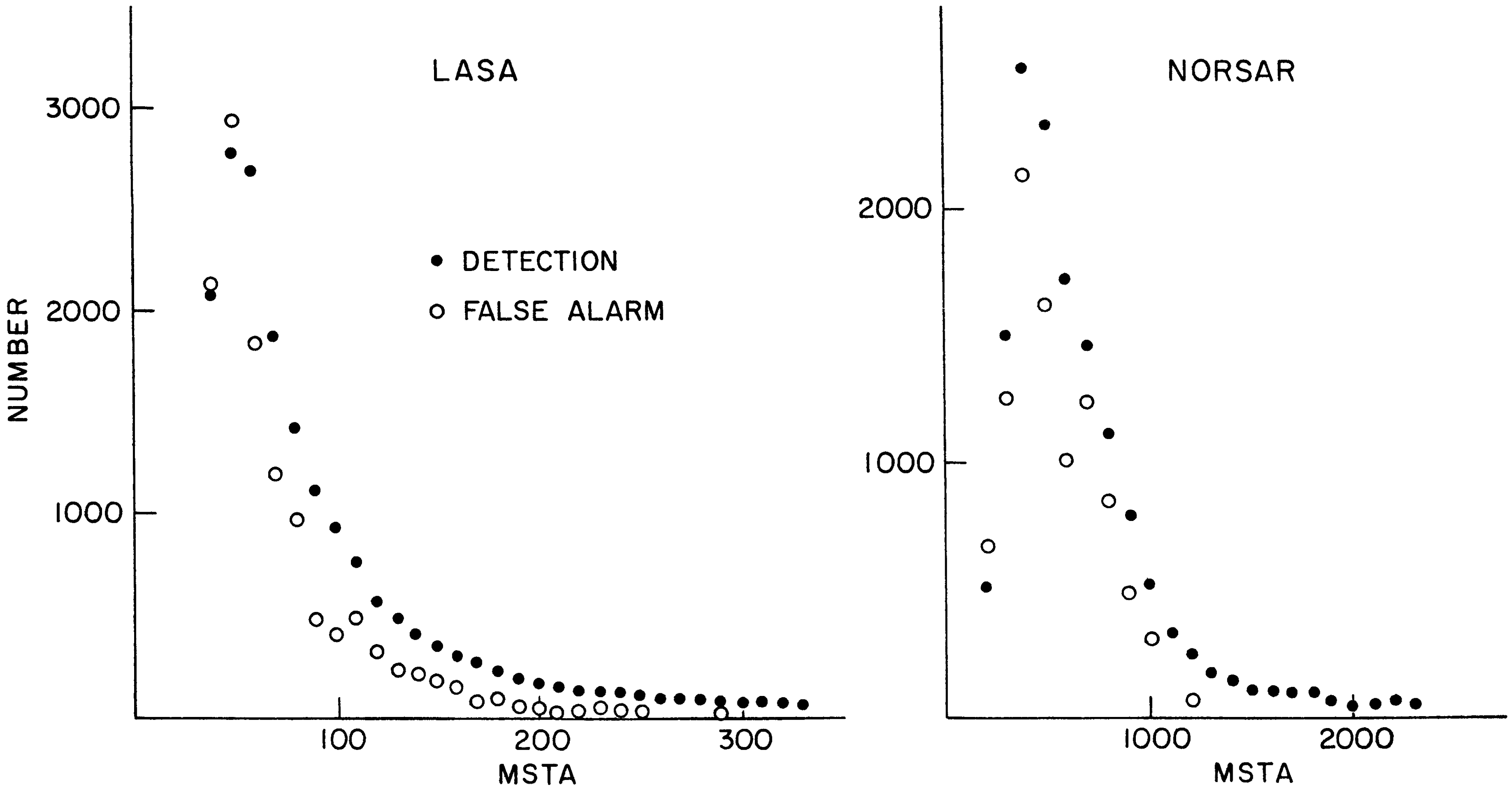
Fig 14



MSTA versus AMPLITUDE LASA

MSTA versus AMPLITUDE NORSAR

Fig 15



LASA
NORSAR
● DETECTION
○ FALSE ALARM
NUMBER
3000
2000
1000
100
200
300
MSTA
2000
1000
1000
2000
MSTA

Fig 16

FALSE ALARM PROBABILITY

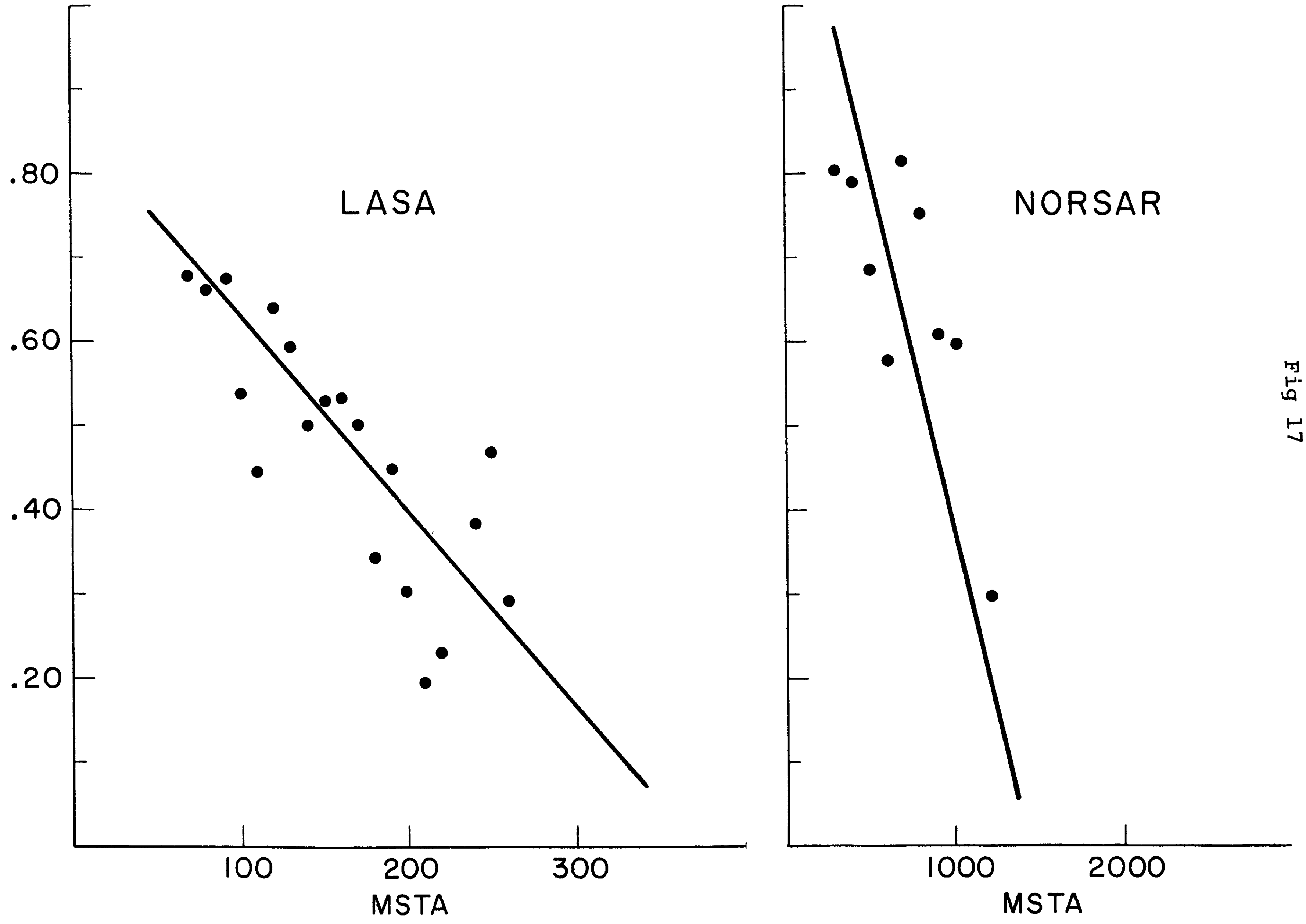


Fig 17

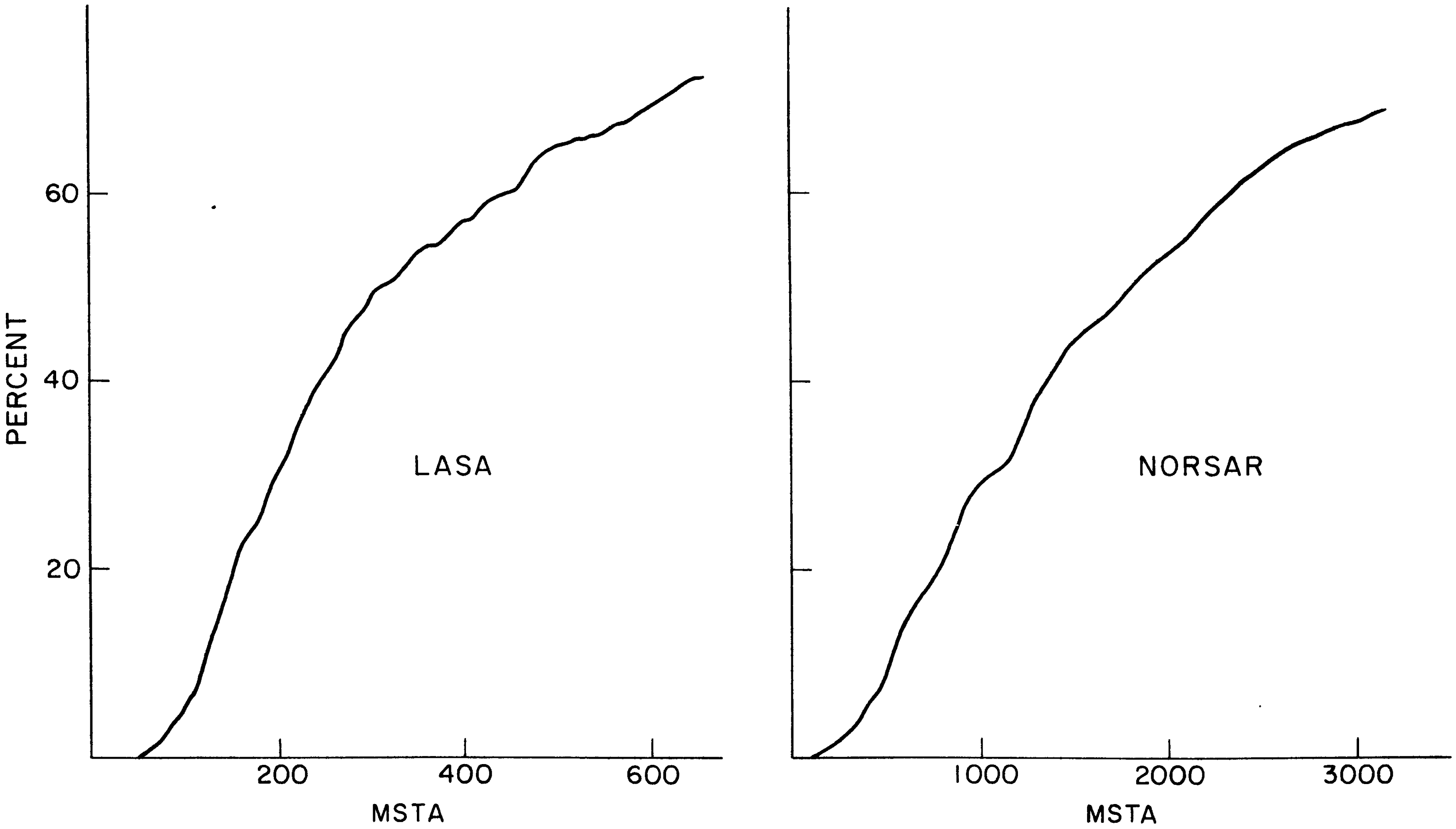
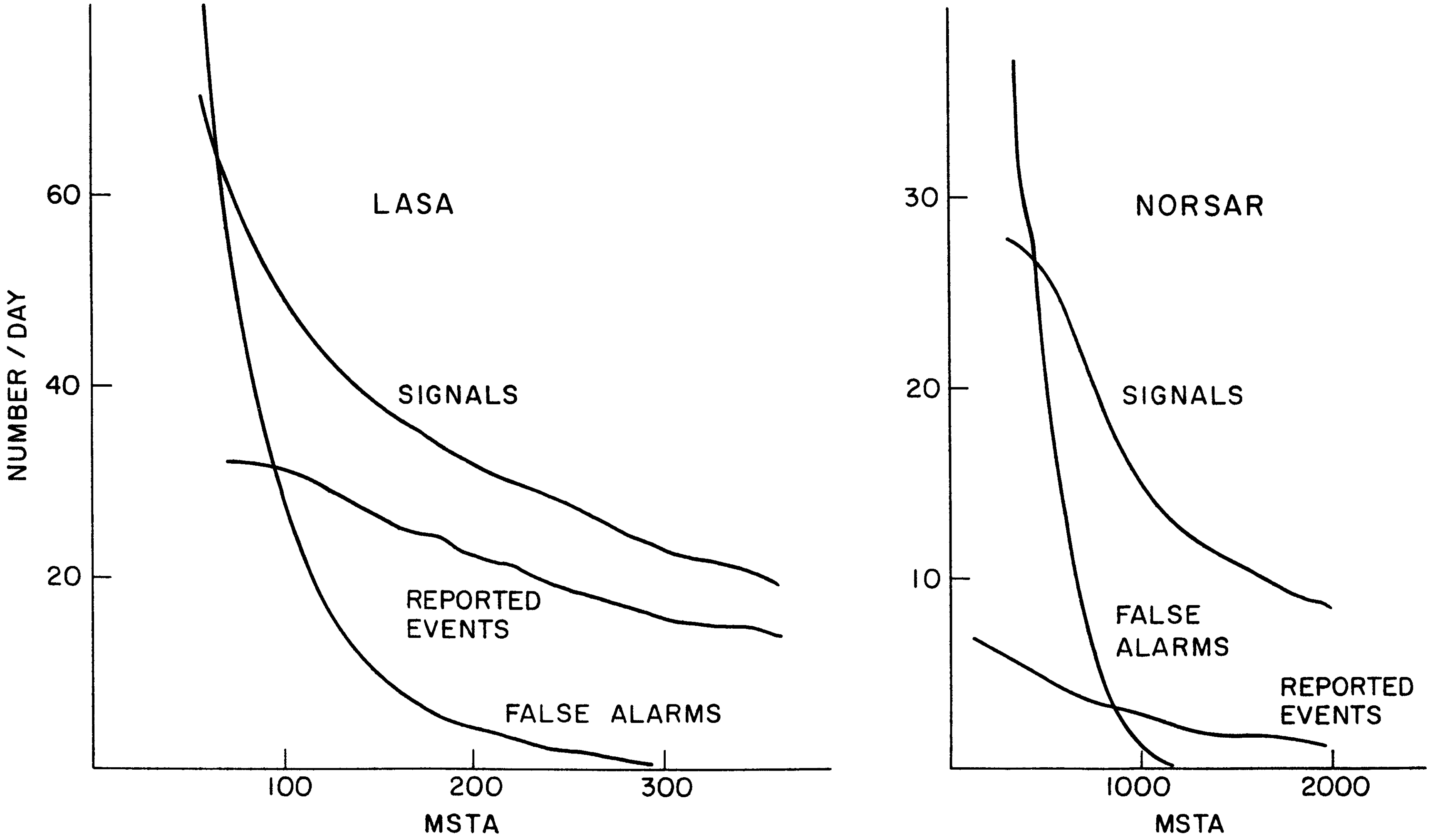

PERCENT
60
40
20
LASA
200
400
600
MSTA
NORSAR
1000
2000
3000
MSTA

Fig 18



LASA
NUMBER / DAY
60
40
20
SIGNALS
REPORTED EVENTS
FALSE ALARMS
100
200
300
MSTA
NORSAR
30
20
10
SIGNALS
FALSE ALARMS
REPORTED EVENTS
1000
2000
MSTA

Fig 19

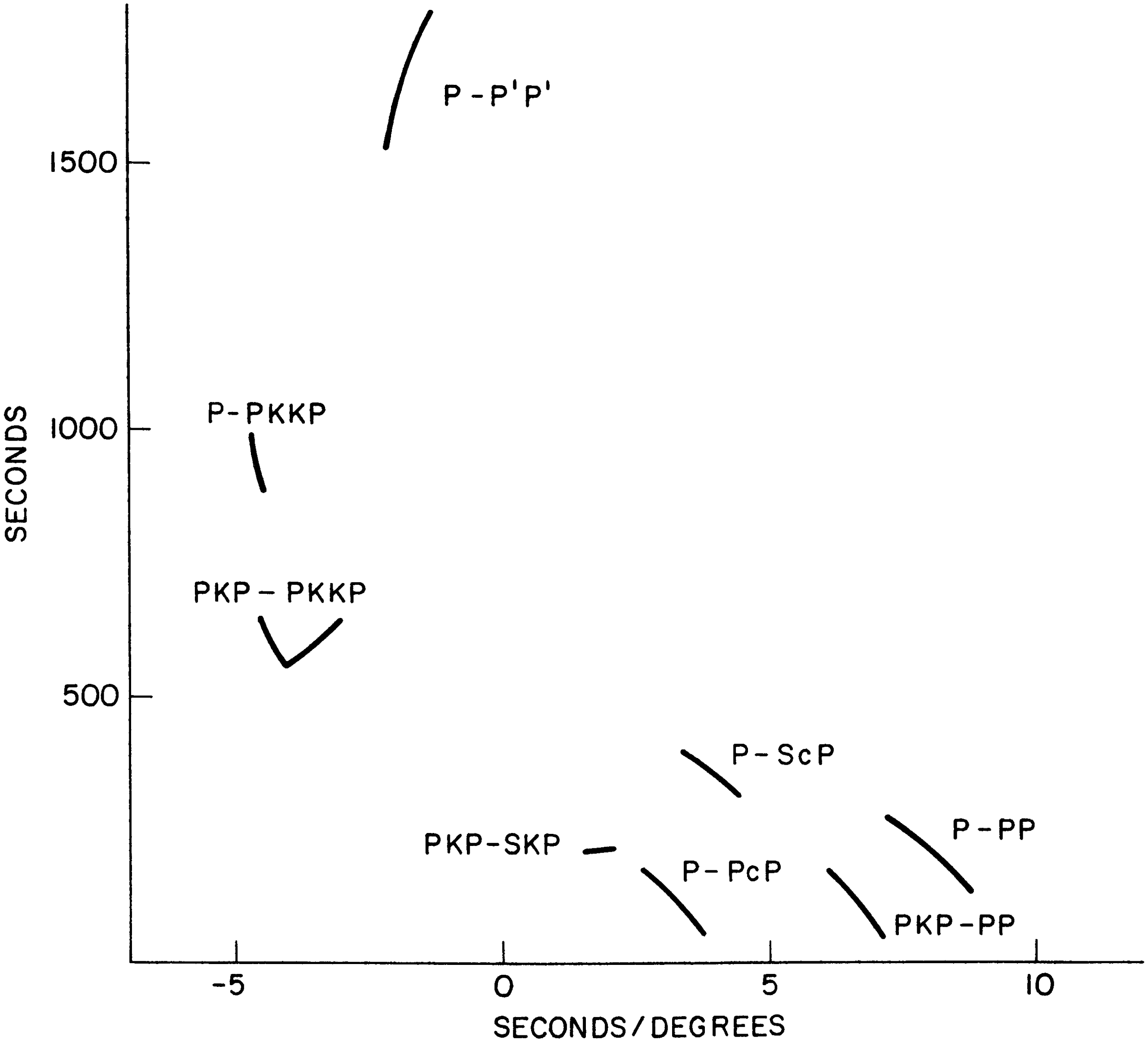

P-P'P'
P-PKKP
PKP-PKKP
P-ScP
P-PP
PKP-SKP
P-PcP
PKP-PP
1500
1000
500
SECONDS
-5
0
5
10
SECONDS/DEGREES

Fig 20

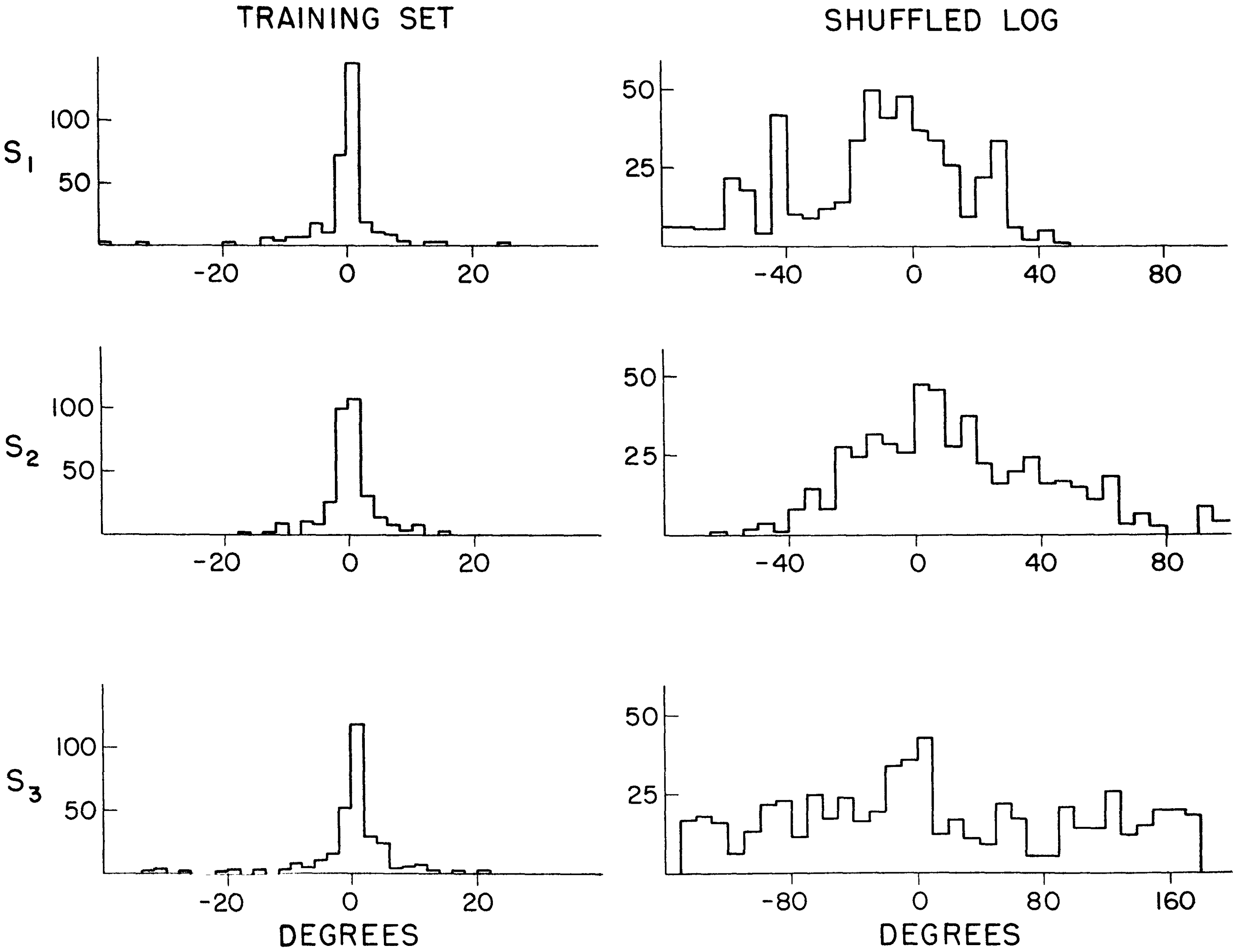

TRAINING SET
SHUFFLED LOG
S_1
S_2
S_3
100
50
25
-20
0
20
-40
40
80
-80
160
DEGREES
DEGREES

Fig 21

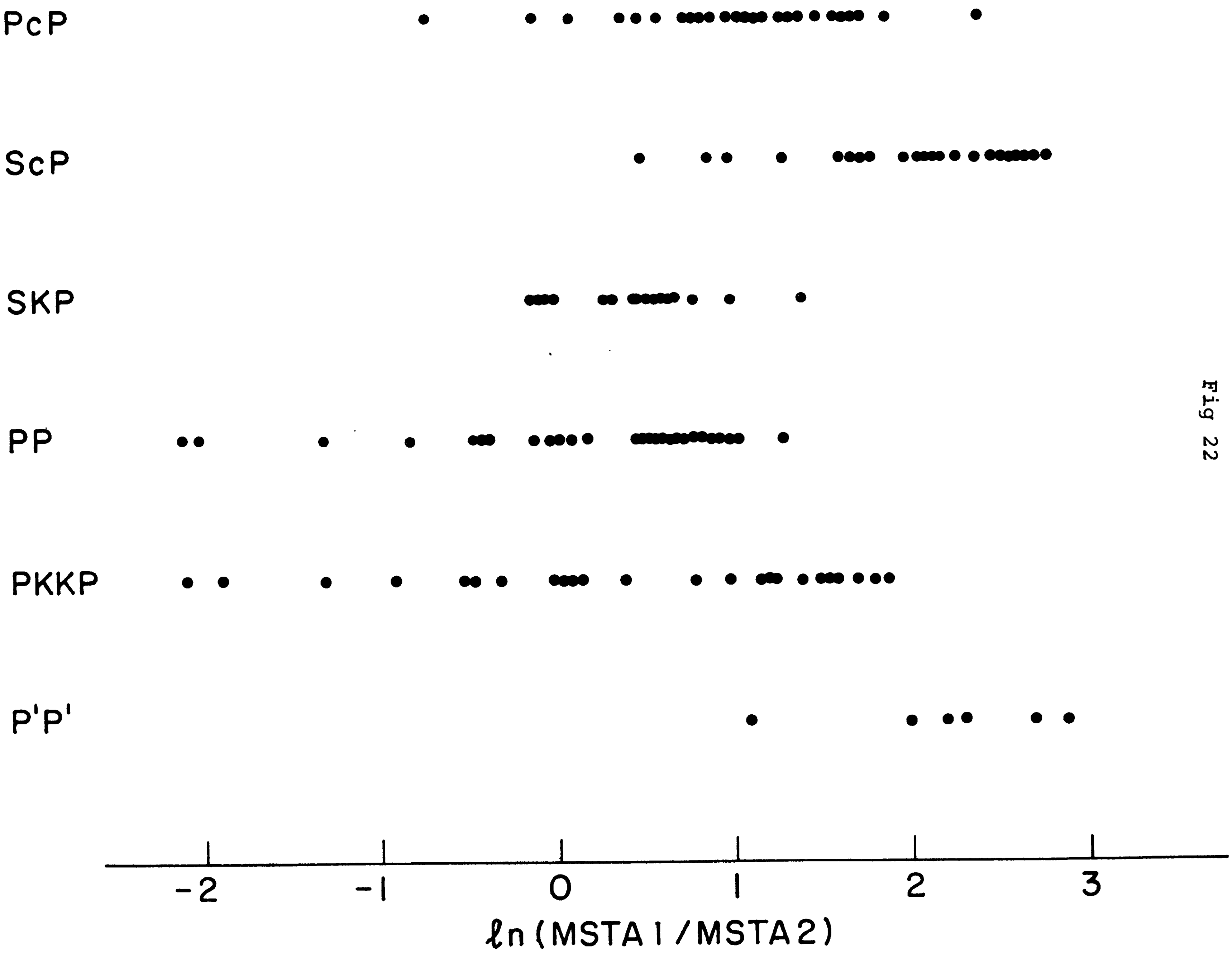


Fig 22

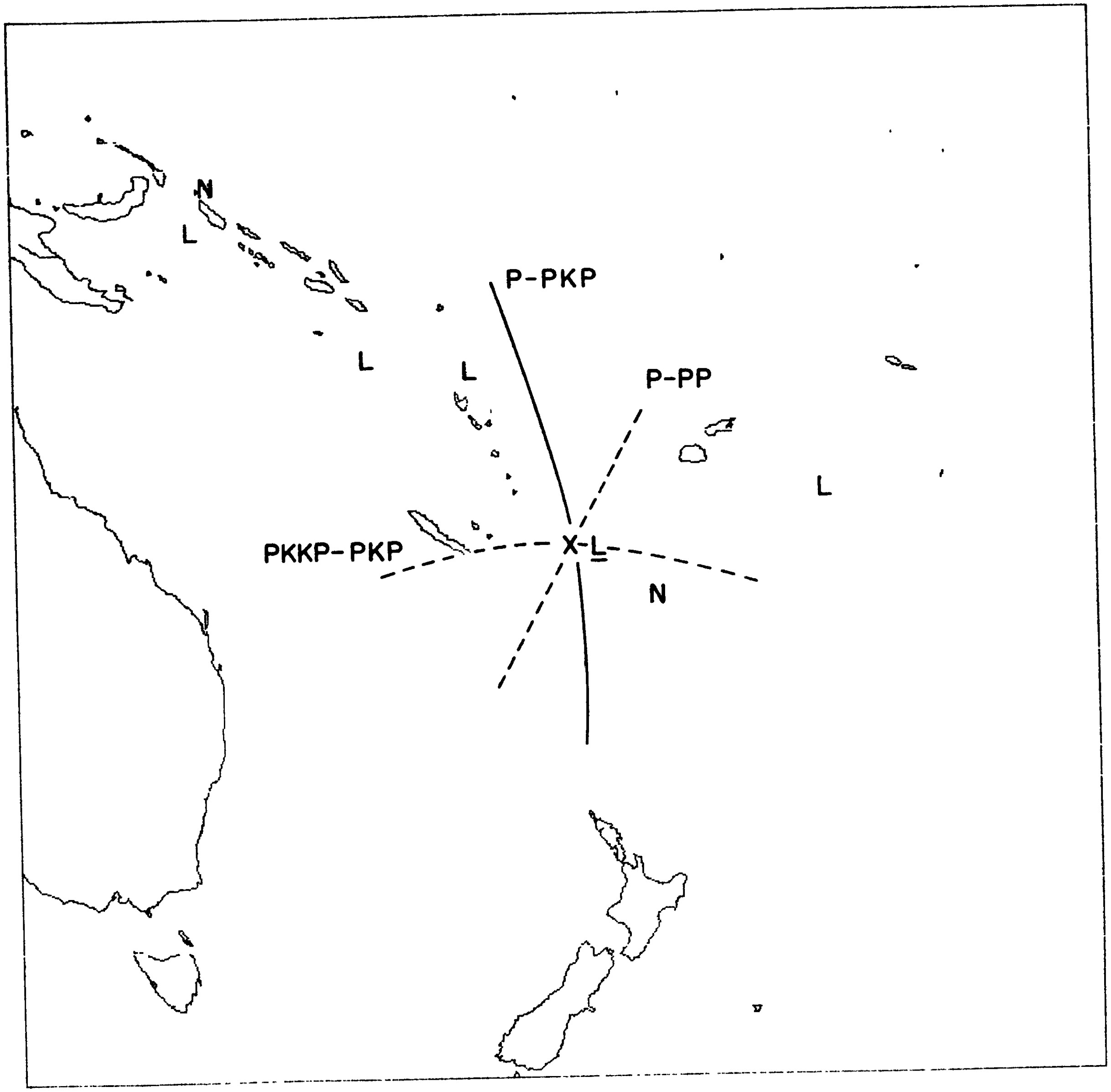


Fig 23

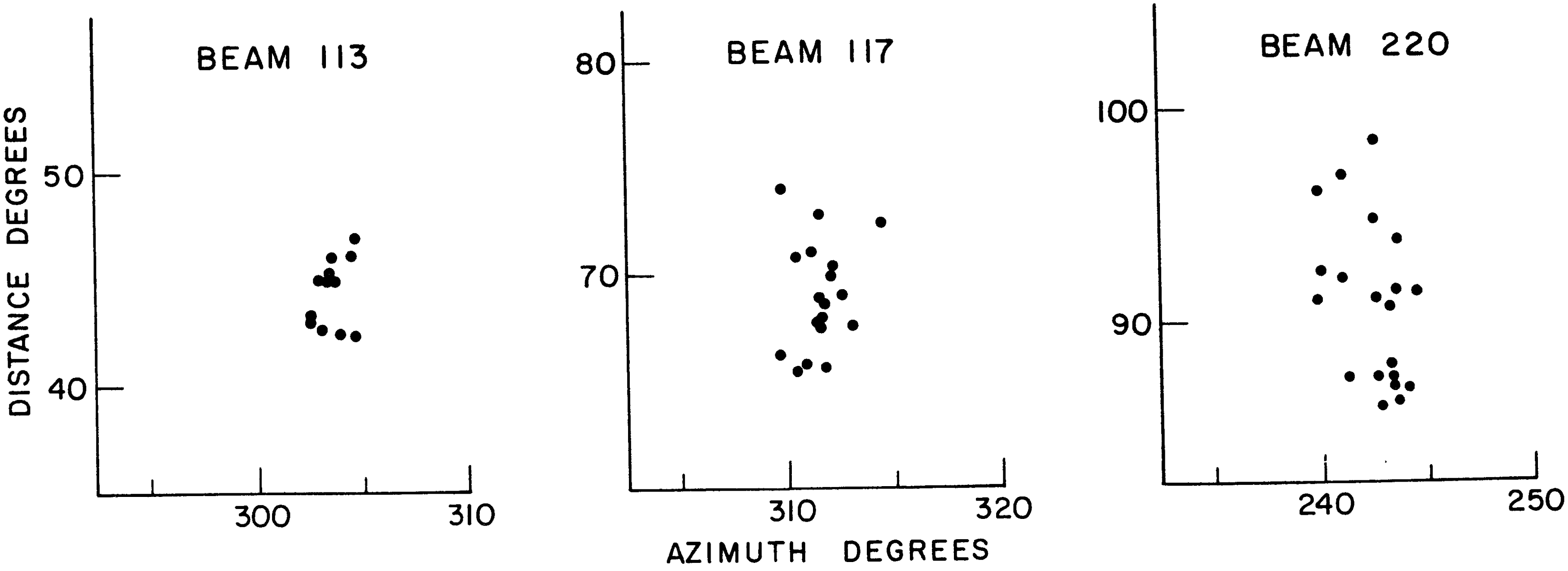
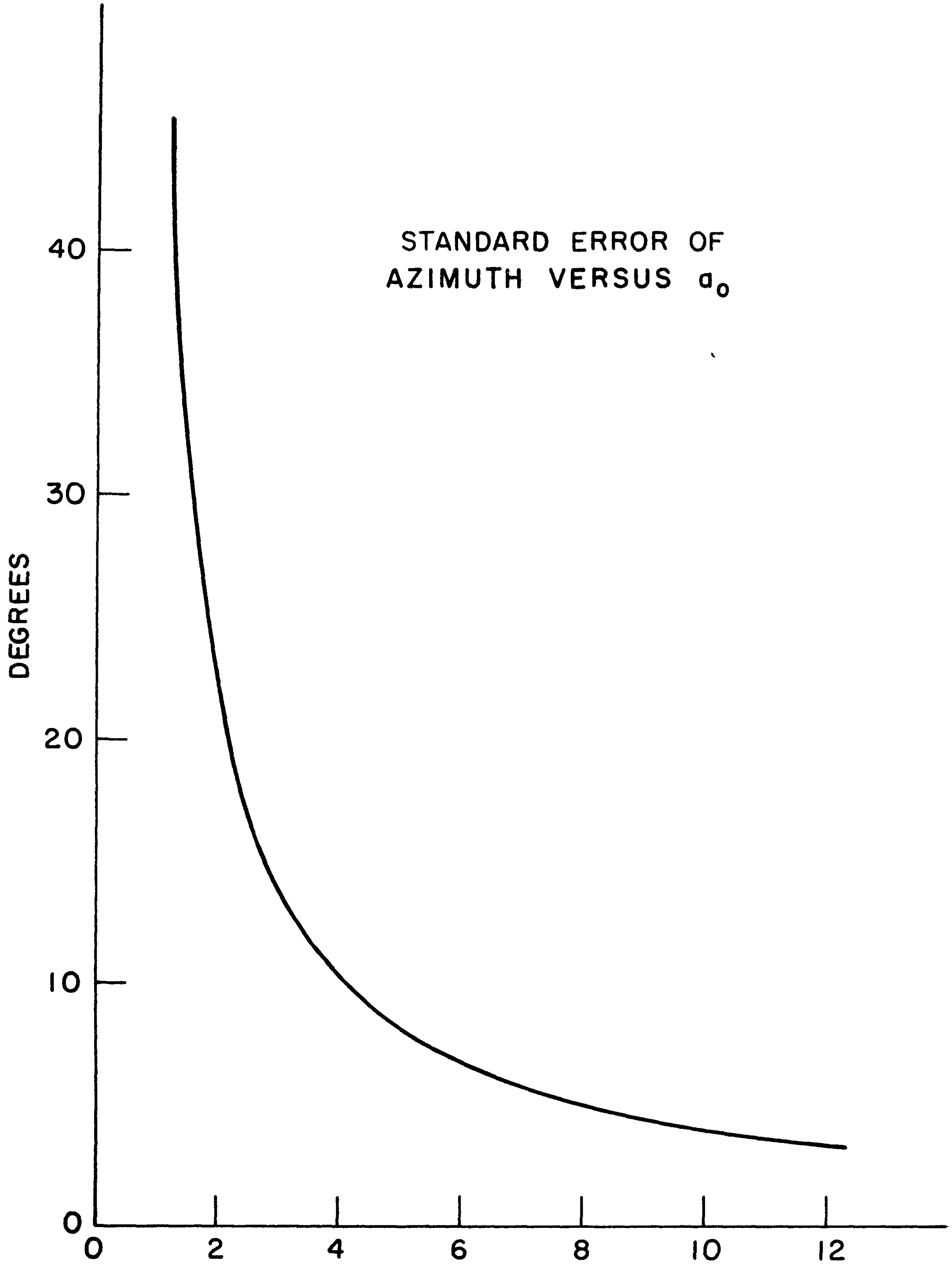

BEAM 113
BEAM 117
BEAM 220
DISTANCE DEGREES
50
40
300
310
80
70
310
320
AZIMUTH DEGREES
100
90
240
250

Fig B-1

Fig B-2


STANDARD ERROR OF
AZIMUTH VERSUS a_0
DEGREES
40
30
20
10
0
0
2
4
6
8
10
12

CUMULATIVE PROBABILITY DISTRIBUTION

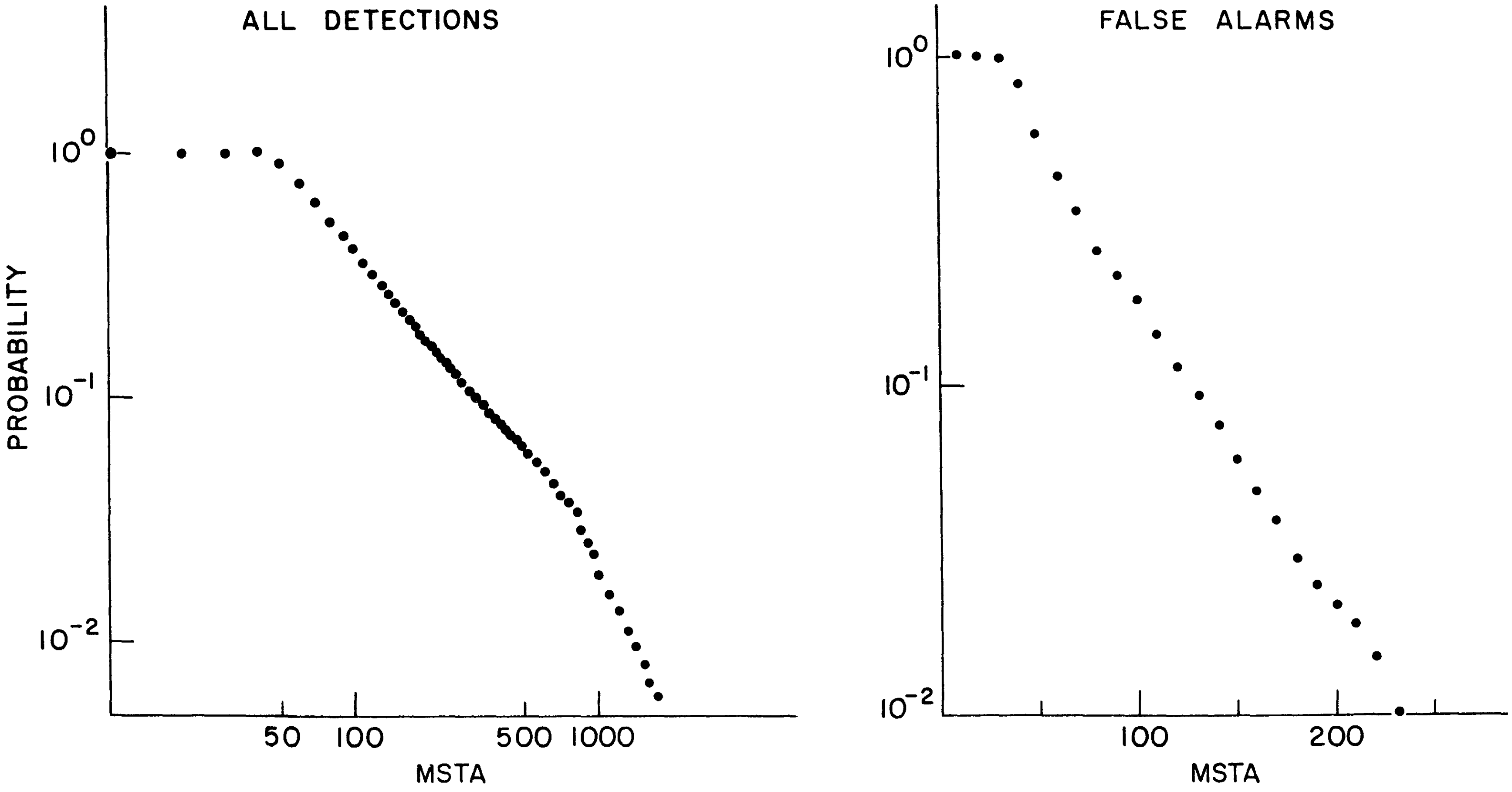


Fig E-1

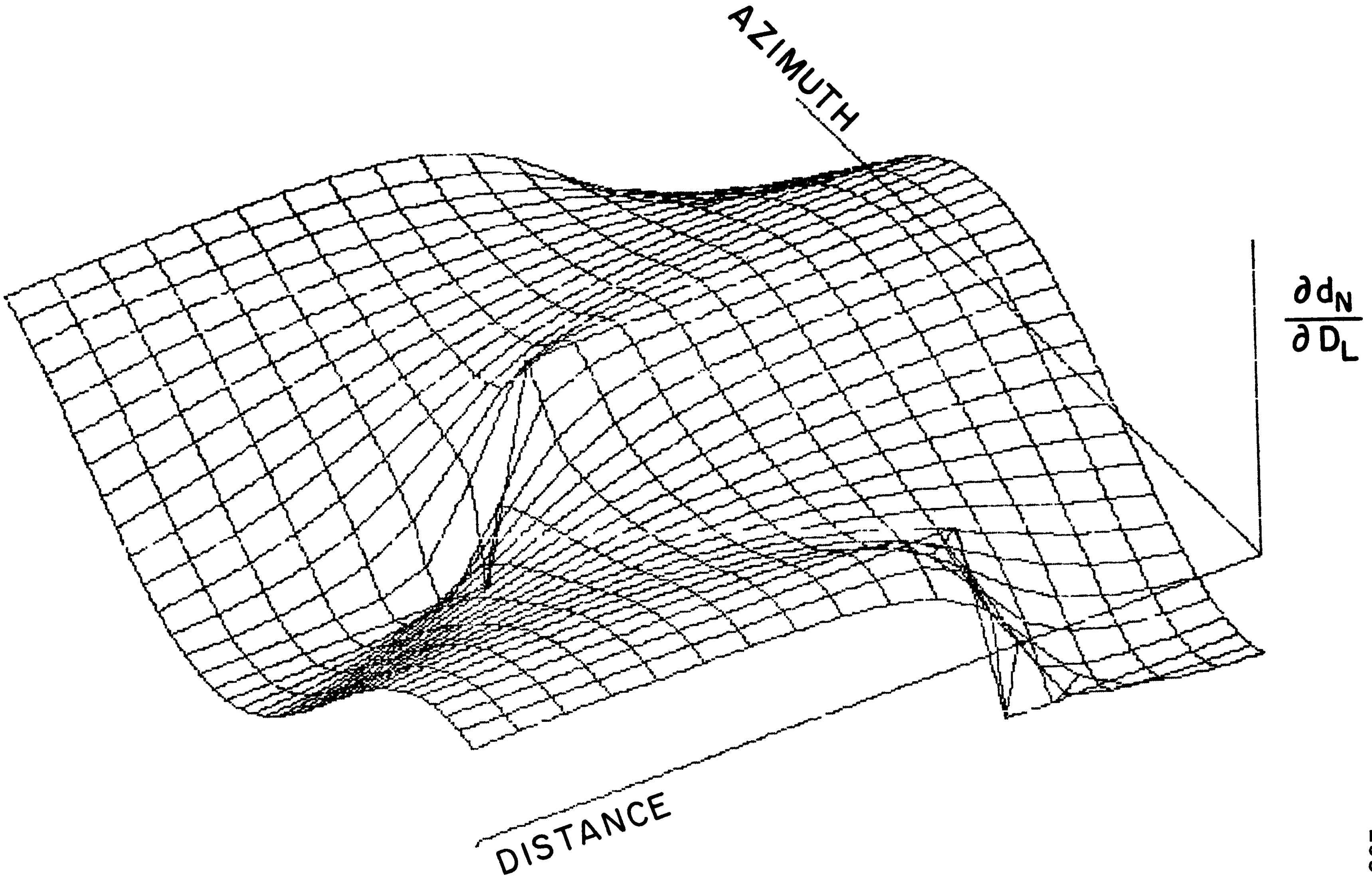


Fig F-1

BIOGRAPHICAL SKETCH

Seymour Shlien was born July 2, 1946, in Montreal, Canada. He did his undergraduate work at McGill University from 1962 to 1968 where he majored in both Geology and Physics. He has been a graduate student in the Department of Earth and Planetary Sciences at M.I.T. since September 1968 and has just completed his Ph.D. thesis on automatic detection and identification of earthquake phases using data from Large Aperture Seismic Arrays.

His field of interest is computer applications to seismology and in particular statistical modeling, decision making, pattern recognition, simulation, and signal processing. He had studied the time occurrences of earthquakes to determine how they differed from a completely random process.

Publications:

Shlien, S. and Toksöz, M.N., "Frequency Magnitude Statistics of Earthquake Occurrences", Earthquake Notes, 41, 5-18, 1970.

Shlien, S. and Toksöz, M.N., "A Cluster Model for Earthquake Occurrences", Bull. Seis. Soc. Am., 60, 1765-1787, 1970.

Shlien, S., Earthquake-Tide Correlation, Geophys. J.R. Astr. Soc., 1972 (in press).