AUTOMATIC PROCESSING OF LOCAL EARTHQUAKE DATA

by

KENNETH ROBERT ANDERSON

B.A., University of Pennsylvania

(1969)

M. S., Massachusets Institute of Technology

(1972)

SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE

DEGREE OF

DOCTOR OF PHILOSOPHY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

August, 1978

Signature of Author...
Department of Earth and Planetary Sciences
September, 1978

Certified by...
Thesis Supervisor

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

AUTOMATIC PROCESSING OF

LOCAL EARTHQUAKE DATA

by

Kenneth Robert Anderson

Submitted to the Department of Earth and Planetary Sciences on August 11, 1978, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

ABSTRACT

A wealth of information about the tectonic process, structure and properties of the earth's crust is being collected daily by the dense network of 148 seismographs in Southern California. However, to fully utilize this information a rapid, accurate, and uniform procedure for analyzing this data is needed.

The Caltech Earthquake Detection and Recording system (CEDAR) is a major step in this direction. It monitors the network and records digital seismograms for later analysis. This thesis studies the process of measuring arrival times and locating earthquakes from CEDAR data in order to develop an objective, reliable computer program to aid seismic analysis.

Analysis of hand picked P wave arrival times from CEDAR shows that beyond about 120 km. distance arrivals tend to be weak and ambiguous. This increases the possibility of picking late arrival times since the first half cycle or so of the onset may be missed or a later arrival might be picked. Thus data from this distance range should be used with care. Travel time residuals are not Gaussianly distributed and each event is likely to have one or more large arrival time errors. The quality assigned to an arrival by an analyst appears to be fairly subjective.

The computer program developed in this thesis locates each event and provides a display of each seismogram in the vicinity of the arrival time. Thus it can provide an interface between the raw CEDAR data and the analyst. For each event, P wave arrival times and the earthquake location are determined by an iterative method. First, arrival times are determined by a simple algorithm applied to each seismogram. A unique aspect of the algorithm is that it objectively estimates arrival time accuracy. The arrivals are used to determine an initial location for the event. Arrivals which were missed or were inconsistent with this location are repicked using a more sensitive algorithm. The earthquake is then relocated. This process may be repeated until a stable set of arrivals is determined.

To make the location procedure resistant to arrival time errors a three step location algorithm is used. Before a location is determined, large errors are removed by a pairwise consistency check. The order in which the arrivals occur across the network is then used to determine a starting epicenter, and the hypocenter is determined by a nonlinear robust regression.

The results of computer processing compare favorably with those of hand processing. Out of the 25 earthquakes analyzed using this procedure, 75 percent of the machine determined epicenters were within 1.2 km. of the corresponding epicenters determined from hand picked arrival times. Of the 400 arrivals picked by both seismologists and the machine, 50 percent agreed to within 0.04 sec. and 75 percent agreed to within 0.12 sec. Only 5 percent of the hand picked arrivals were missed by the machine. These were all weak, low quality arrivals. The machine picked 25 percent more arrivals than the seismolgist. These arrivals were all very weak, ambiguous arrivals simply ignored by the seismologist during the initial viewing of the data. The accuracy assigned to each pick agrees well with that assigned by the seismologist and provides a 90 percent confidence interval for the arrival time.

Thus much of the first stage processing of CEDAR data can be done automatically. Since the computer provides the seismolgist with much more information than he would normally have the first time he views the data, it is much easier to produce an accurate and objective data base which is so important in seismic research.

Thesis Supervisor: Keiiti Aki, Professor of Geophysics

ACKNOWLEDGEMENTS

The following people, animals, and machines have contributed in some way to this thesis or its author's well being: K Aki, Rex Allen, Anne Anderson, Tore Anderson, Ann Blanchard, Michel Bouchon, Raymon Brown, Sara Brydges, Herman Chernoff, Bernard Chouet, Anton Dainty, David Edelman, Bill Ellsworth, Neil Goins, B.K.P. Horn, Kathy Huber, Darya Huber-Anderson, Carl Johnson, Dick Lacoss, Larry Lande, Willie Lee, Alan Lindh, Tomås Lozano-Perez, Eleo Madariaga, Raul Madariaga, Veruschka von Maitau, MIT-AI, MULTICS, Peter Neilson, Paul Reasenberg, Robin, Randy Richardson, Frank Press, Riverwood's Chryse Planitia, Jack Sax, Larry Sargent, Jim Scheimer, Ken Schroder, Tony Shakal, Bob Sheppard, Judy Stein, Pete Stevenson, Sam Stewart, Nafi Toksöz, UNIX, Pete Ward, Roy Welsch, Ralph Wiggins, and George Zandt. Unfortunately, time does not permit more fitting and personal acknowledgements. I would like to thank the Lincoln Laboratory, Applied Seismology Group, for use of its facilities. This work was supported by the United States Geological Survey under contracts 14-08-001-G-339 and 14-08-0001-16761.

TABLE OF CONTENTS

1. INTRODUCTION

The microearthquake data collected by a dense network of seismographs such as the central California array operated by NCER (National Center for Earthquake Research), U.S.G.S.or the joint Caltech - U.S.G.S network in Southern California may contain a wealth of information about the tectonic process and structure of the earth's crust under the array.

Microearthquake data is currently being used in earthquake predicition, locating and monitoring geothermal areas and three dimensional seismology. The use of microearthquakes in earthquake predicition is based on the idea that large earthquakes share the same tectonic causes as the numerous small ones occurring in the same general area (Aki, 1968). Detailed fault zone structure can be determined from the location of microearthquakes (Eaton _et al_., 1970). Tectonic stress indicators such as fault plane solutions and stress drop can be determined from first motion studies and spectra (McNally & McEvilly, 1977; Aki, 1967, 1968; Brune, 1970). Precursory variations in a large number of geophysical parameters are currently being studied. Variations in Vp/Vs or P wave travel time residuals (Semenov, 1969; Aggarwal _et al_., 1973; Robinson, Wesson and Ellsworth, 1974), epicenter recurrence patterns, anomalous seismicity (Kerr, 1978), and temporal changes in fault plane solutions show promise as predictive indicators. The short term indicators such as anomolous seismicity and changes in fault plane solutions may be particularly effective in an earthquake prediction program (Lindh, Fuis, and Mantis, 1978; Kerr, 1978).

Microearthquakes can be used in three dimensional seismology to determine structural details around an active fault zone and map geothermal areas and magma chambers. In three dimensional seismology, the arrival times from a large number of stations and events are used to determine the fine scale structure of a small area. The ultimate resolution of three dimensional seismology depends on the wavelength of the first arrival, which may be about a few hundred meters for microearthquakes. Experiments at the NCER array in California give a resolution of about 5 km. using microearthquake data (Aki and Lee, 1975). In these experiments, the resolution is limited by the number of stations. A higher density of stations would increase the resolution.

The full exploitation of this data for the purpose of earthquake prediction has been hampered by a lack of a strict, uniform procedure for analyzing microearthquake data. For example, the bias and error in elementary measurements such as picking the time of first arrivals may vary from time to time because of a change in personnel hired for the work. Workers in statistical seismology are often dismayed by the non-uniformity of the data set in time, when they are studying epicenter migrations, sympathetic occurrence of earthquakes at different places, foreshocks, the changes in b-value, fault plane solutions, and the Wadati diagram. Unfortunately there is evidence that at least one anomalous P-delay attributed to dilatancy may be due in part to subjective bias of the personnel who read the records (Lindh, Lockner, and Lee, 1978).

It is impossible to avoid subjective bias in reading seismograms; some people tend to pick arrivals earlier than the real one, others tend to pick later ones (Freedman, 1966a, 1966b, 1968). It is also impossible to get any objective estimate of the errors involved (see Section 6.1).

This is important in first motion studies. Aki (1976) has shown that often the wrong sense of first motion is picked if the signal to noise ratio is lower than a critical value. Pearce and Barley (1977) come to the same conclusion.

The California seismic networks record an enormous amount of data annually. For example, in one year the Southern California network recorded over 260,000 seismograms from over 7000 events, and at least as many seismograms can be expected from the Central California network each year. Even if seismograms can be processed manually at this rate, earthquake swarms and aftershock sequences can increase the seismicity level of an area by an order of magnitude. Rapid, objective analysis of such a volume of data could be crucial for earthquake prediction (Lindh, Fuis, and Mantis, 1978). The bulletin which is a formal report of the routine analysis of the NCER data is presently two and one-half years behind schedule (W. H. K. Lee, personal communication, 1976).

Recently, the rate at which seismic data can be analyzed has been vastly increased by the use of the California Institute of Technology Earthquake Detection and Recording (CEDAR) system. CEDAR is a real time computer system which detects earthquakes on the Southern California array and records the seismograms for later processing. The final output of CEDAR is a data base which includes earthquake locations, P and S wave arrival times, and the seismograms. Thus for the first time, CEDAR provides all of the raw materials that a seismolgist needs in a format that is easily manipulated by computer. What is now needed are the computational tools to aid seismologists in the interpretation of the CEDAR data.

In this thesis, the process of how seismologists measure arrival times and locate earthquakes is studied in order to develop an objective and reliable computer program to aid in seismic analysis of the CEDAR data.

The main points of the thesis may be summarized as follows. There is always ambiguity associated with measuring first arrival time from seismograms whether it is done by a seismologist or by a machine since the seismic signals are of unknown shape and are contaminated with noise. This ambiguity increases with distance from the epicenter because of the structure and attenuation of the earth. For example, beyond 100 km. distance the first arrival becomes very weak and can easily be confused with the much stronger Moho reflection.

In order to reduce the effect of this ambiguity, information from many sensors must be used to constrain arrival times on individual sensors. Since one of the most important constraints is the location of the event, the processes of picking arrivals and locating the event should be combined in an iterative fashion.

First arrival times are determined by an algorithm applied to each seismogram. A unique aspect of the algorithm is that it objectively estimates arrival time accuracy. The arrivals are used to determine an initial location for the event. Arrivals which were missed or were inconsistent with this location are repicked using more appropriate parameter settings. The earthquake is then relocated. This process may be repeated until no arrivals need to be repicked.

The power of the iterative algorithm comes from the feedback provided by the locating and advising stage. It is only the locating stage which has enough information to evaluate the performance of the picking stage. The success of the algorithm will depend critically on how good the initial picks are, how the locator handles bad arrivals, how good the advice is, and how well the picker can utilize advice.

1.1. AN EXAMPLE

An example of the automatic processing of CEDAR data is shown in Exhibits 1.2 and 1.1. Exhibit 1.1 shows five seconds of seismic data, two seconds are before the machine pick and three seconds are after. The final machine pick and its confidence limits is shown by the three dotted lines and the arrival time predicted from the location is shown by the long dashed lines. The location parameters and a description of each arrival are shown in Exhibit 1.2. If a hand picked arrival time is available, the difference between the machine and human arrival time (M-H) and the quality of the hand pick is shown. For the eight arrivals which were repicked, a comment about the first pick is given.

Two iterations of the algorithm were applied to this event. On the first iteration, thirty four seismograms were inspected and the picking algorithm picked arrivals on fifteen of them. The locating algorithm identified arrivals from two distant stations as inconsistent. One arrival is an S arrival, and the other appears to be from a different event local to that station. The advising stage advised that the nineteen seismograms which had no picks and three stations which had bad picks be repicked with adjusted parameter settings. On the second iteration, five more very weak arrivals (marked "none") were picked, although only three arrivals have small residuals. Of the three seismograms that were picked twice, only one arrival time was improved.

Considering that the event is quite small (coda magnitude 1.5), the algorithm does quite well. Of the fifteen arrivals, only five have large travel time residuals, and only three are significantly different from their corresponding hand picks. Of the eight arrivals repicked, six of the arrival times were improved, and two did not change.

1.2. OVERVIEW OF THESIS

The next chapter will begin with a description of the CEDAR system. Since CEDAR had only recently begun operation when the data used in this thesis was collected, the data contained a number of problems which are discussed. A summary of the operational characteristics of each station is determined from CEDAR's detection log.

Chapter 3 summarizes work by Freedman and others on how seismologists determine arrival time. There is often considerable subjective bias in how arrival times are picked and the quality assigned to them. This bias does not necessarily decrease with the skill of the seismologist! CEDAR hand picked arrival times will be looked at in light of Freedman's results. A statistical model of human picking errors will be built to compare with automatic picking errors.

Chapter 4 begins with a discussion of some of the problems which complicate the design of picking algorithms. Using information from other sensors to clarify the picking process is stressed. The picking algorithm used is presented in Section 4.1. It is similar to one used in the Computer Detection System, CDS, which monitors the U.S.G.S. Central California Seismological Network (Stewart, 1977). The algorithm for determining arrival time accuracy is presented in Section 4.2.

Chapter 5 will be concerned with robust earthquake location. Locating an earthquake by Geiger's method involves the solution of a set of non-linear equations. Usually, an iterative procedure is used which starts from a trial epicenter. Each step in the iteration requires the solution of a linear least-squares problem (Buland, 1976; Lee and Lahr, 1975). Reliable locations result from this procedure only if the location method is robust; i.e., the location is not seriously degraded by a few large errors among the arrival times. To make the procedure more robust, the equations for the next iteration may be weighted in relation to their corresponding travel time residuals (Lee and Lahr, 1975; Bolt, 1976).

This procedure works well if the initial location is relatively close to the true epicenter. In common practice, the location of the station reporting the earliest arrival is used as the initial location. Unfortunately, in the case of the CEDAR data studied, the median distance from the first station reporting the event to the epicenter was 11 km., and using this location to initialize the algorithm could add a second or more to the residuals. The actual outliers could then be hidden in these inflated residuals. On the other hand, using the unweighted least-squares location to initialize the algorithm also does not guarantee a robust location (Howard Patton, Personal Communication, 1977; Andrews, 1974). This is because the least-squares method tends to reduce the larger residuals more than it reduces the smaller ones (Claerbout and Muir, 1973; Andrews _et_ _al_. 1972).

To circumvent these difficulties, a new method which uses the order in which arrivals occur at different stations has been developed to determine a trial epicenter. Since only the arrival order is used and not the actual arrival times, grossly inconsistent arrivals will not affect the location. One interesting feature of the method is that it does not depend on a crustal velocity model. The results of several location methods are compared in Section 5.3.2 and the effect of arrival time error on least squares and robust location for one event is considered in Section 5.3.3.

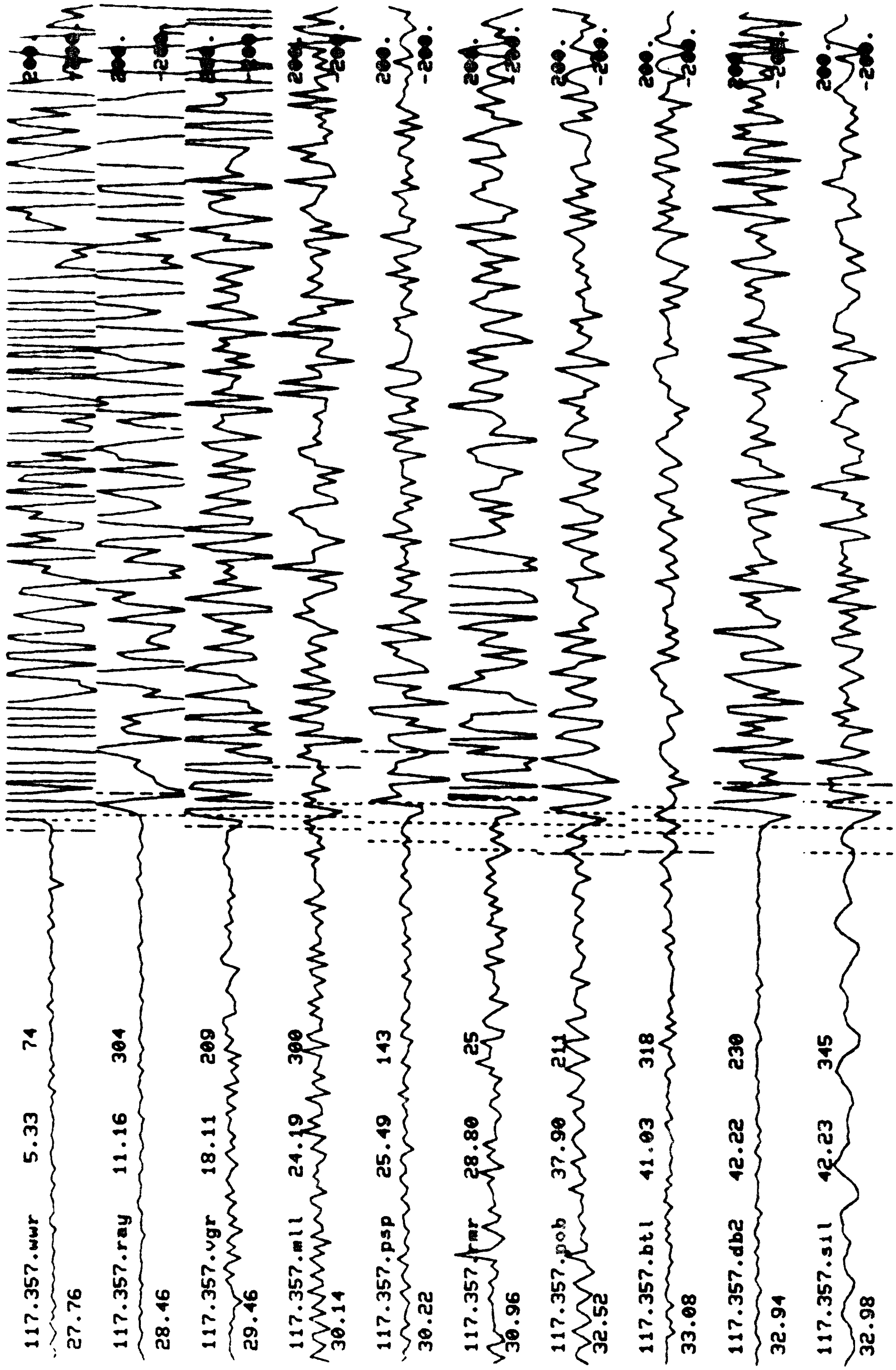
In Section 5.4, arrival picking and locating is combined into an iterative algorithm, and human and machine processing is compared in Chapter 6. Some evidence for human bias is presented.

The thesis concludes with recomendations to improve both the algorithm and CEDAR and a discussion of the future of automatic seismic analysis.

1.3. EXHIBITS

1.1. Suite of machine picked first arrivals. For each seismogram, 5 seconds of data, 2 before and 3 after the machine pick are shown. (Time increases from left to right.) Each seismogram is annotated as follows: From left to right, the three vertical dotted lines indicate $t-2\sigma_t$, t, and $t+2\sigma_t$, where t is the machine arrival time and σ_t is the machine determined standard error. $\pm 2\sigma_t$ provides a 90% confidence limit for the arrival (see Section 6.2). In the upper left hand corner is the event number, the station name and the distance and azimuth of the station from the event. The time of the left most data point displayed and the time interval between tickmarks, if any, (in seconds) is shown in the lower left hand corner. The maximum and minimum amplitude displayed is shown in the upper and lower left hand corner.

1.2. Location parameters and arrival times for event 117.357. For each arrival, the residual is the machine picked arrival time minus the theoretical arrival time; the weight is the product of the standard error assigned by the picking algorithm and the robust weight determined by the location algorithm. When corresponding hand picks are available, M-H is the machine arrival time minus the hand picked arrival time; and the quality is the hand assigned quality (see Section 3.2).



117.357.wwr 5.33 74
27.76
117.357.ray 11.16 304
28.46
117.357.vgr 18.11 209
29.46
117.357.mll 24.19 300
30.14
117.357.psp 25.49 143
30.22
117.357.tmr 28.80 25
30.96
117.357.pob 37.90 211
32.52
117.357.btl 41.03 318
33.08
117.357.db2 42.22 230
32.94
117.357.sil 42.23 345
32.98

EXHIBIT 1.1

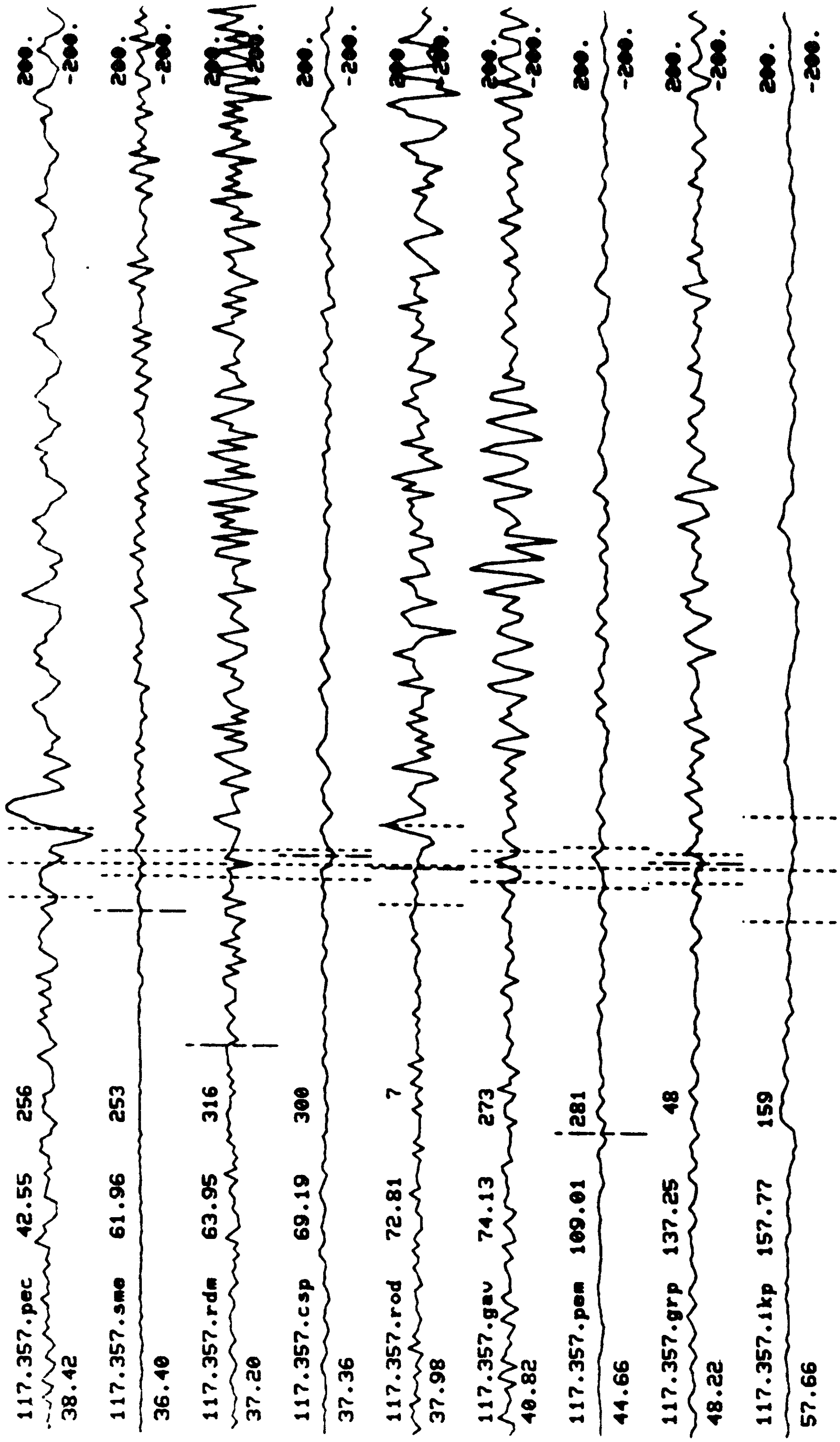


EXHIBIT 1.1 CONTINUED

EXHIBIT 1.2

LOCATION PARAMETERS AND ARRIVALS FOR EVENT 117.357

ORIGIN TIME	LATITUDE	LONGITUDE	DEPTH
28.24	33.979	-116.712	6.86

STATION	ARRIVAL TIME (SEC.)	STANDARD ERROR (SEC.)	RESIDUAL (SEC.)	WEIGHT	DISTANCE (KM.)	AZIMUTH	M-H (SEC.)	QUALITY	FIRST PICK
wwr	29.76	0.001	0.04	99	5.33	74	0.02	100	
ray	30.48	0.009	-0.08	98	11.16	304	0.0	100	
vgr	31.48	0.009	0.04	99	18.11	209	0.06	75	
mll	32.16	0.024	-0.18	36	24.19	300	0.03	50	NONE
psp	32.22	0.036	-0.26	22	25.49	143	-.02	75	
rmr	32.96	0.049	-0.09	20	28.80	25	-.02	75	
pob	34.52	0.022	0.11	44	37.90	211	0.0	75	
btl	35.10	0.024	0.12	40	41.03	318			
db2	34.94	0.006	-0.16	91	42.22	230	0.0	100	
sil	34.98	0.048	-0.16	19	42.23	345	-.01	100	
pec	40.42	0.061	5.26	0	42.55	256			NONE
sme	38.40	0.022	0.17	40	61.96	253	0.44	50	
rdm	39.20	0.024	0.64	5	63.95	316	0.42	0	SAME
csp	39.36	0.026	-0.03	38	69.19	300	0.02	50	NONE
rod	39.98	0.070	0.01	14	72.81	7			
gav	42.82	0.027	2.66	0	74.13	273			NONE
pem	46.66	0.036	0.94	0	109.01	281	1.1	25	S
grp	50.22	0.026	-0.02	38	137.25	48			NONE
ikp	59.66	0.092	6.14	0	157.77	159			SAME

2. CEDAR: CALTECH EARTHQUAKE DETECTION AND RECORDING SYSTEM

The California Institute of Technology Earthquake Detection and Recording (CEDAR) system records digital seismograms from 120 stations of the joint USGS-CIT seismic network in Southern California. CEDAR detects events and records the seismograms on magnetic tape in real time. Arrival times are then measured and the earthquakes are located using a non-real time interactive computer program (Carl Johnson, personal communication, 1978). The final product of CEDAR is a magnetic tape containing the following data for each earthquake:

1. event parameters; e.g., location, origin time, magnitude,

2. P and S wave arrivals, and

3. the corresponding seismograms.

Thus CEDAR provides for the first time a data base containing three levels of seismic information which can be easily manipulated by computer. It is a valuable source of information about tectonic processes, structure and properties of the earth's crust in Southern California. For example, CEDAR has been in operation since January, 1977 and in its first year of operation recorded over 7000 earthquakes, 260,000 seismograms, and 150,000 arrival times.

One week of seismic data recorded by CEDAR in April 1977 was selected for study. It consisted of over 2500 seismograms, nearly 1300 hand picked P wave arrivals, and nearly 600 hand picked S wave arrivals from 63 seismic events. The stations in the CEDAR network are shown in Exhibit 2.1 and the earthquake epicenters are shown in Exhibit 2.2.

Since CEDAR had been in operation for only four months when this data was collected, several problems with this data became immediately apparent:

1. CEDAR failed to record the onsets of some of the earliest seismograms,

2. CEDAR stopped recording before the end of the coda on some seismograms, and

3. many seismograms were contaminated by aliasing as no anti-alias filters are presently used in the system.

These problems will be discussed briefly here since they affect the automatic processing applied to the data as well as the usefulness of the data itself.

The first two of these problems can presumably be solved by adjusting the detection parameters of CEDAR to provide more data at each end of the seismogram. The third problem is disconserting since it significantly reduces the usefulness of the CEDAR data base. Many of the seismograms contain portions which appear to have significant energy at the Nyquist frequency. [1] For example, Exhibit 2.3 shows several first arrivals that have been contaminated by aliasing. The tick marks are 0.2 seconds apart. Aliasing is indicated by the presence of oscillations which have only one point per half cycle; i.e., oscillations at 25 Hz. Higher frequencies are folded down into the digitized data and contaminate the spectrum below 25 Hz.

---

[1] Since the data is sampled at 50 Hz., the Nyquist frequency is 25 Hz.

Since most of the instruments of the Southern California array have short period responses, the aliasing problem affects most of the array. The seismic data that has been examined indicates that as much as fifty percent of the seismograms from an event may contain aliased portions. The array contains approximately fifteen Benioff instruments which have a velocity response corner at 1 Hz. Data from these stations appear to be relatively unaffected by the aliasing problem. These stations are well distributed over the array and, taken together, can be used as a sub-array to process teleseismic data (see Exhibit 2.4). However, the Benioff sub-array is not dense enough to provide adequate coverage for local events.

Earth attenuation provides some help by removing higher frequencies from distant seismograms. The attenuation effect appears to be significant only for distances greater than 60 km. from the epicenter. Thus if we are interested in doing anything besides picking first arrivals, aliasing forces us to use either a sparse network or, at best, the worst half of the data.

There are three obvious solutions to the aliasing problem:

1. increase the sampling rate of CEDAR,

2. apply analog anti-alias filters to each seismic channel before the data is digitized, or

3. apply digital low pass filters to the existing aliased digital seismograms as a stop gap measure.

The first solution is attractive since it would not reduce the bandwidth of the CEDAR system. However, since CEDAR is currently working near full capacity, this would require an additional CEDAR system to handle the increased sampling rate. A modular, extendable system might be able to handle such alteration more easily, but this would require an additional hardware and software effort. Using anti-alias filters seems to be the most practical solution to the problem. CIT is currently experimenting with several anti-alias filters (Carl Johnson, personal communication, 1978). The filters will have a 5 Hz. corner and roll off at 12 db/octave above the corner.

Although installing anti-alias filters would reduce the resulting bandwidth of the instruments, the frequency band below 5 Hz. should still contain a great deal of useful information. This is because much of the high frequency energy is due to scattered waves. In fact, the reduced bandwidth may be quite beneficial. Carl Johnson has already found that it is easier to pick first arrivals from filtered data. Although the resolution of arrival times may be slightly worse (see Section 4.2), the number of errors due to ambiguous arrivals should be reduced. This may have a significant effect on the overall quality of the CEDAR arrival times (Freedman, 1966a, 1966b, 1968).

The anti-alias filters must be chosen with some care since, even though the amplitude response (fall off) of the anti-alias filters is outside the band of interest, their phase response will usually affect data in the band of interest. Furthermore, the group delay introduced by the phase response will bias arrival times from different instruments. For example, McCowan and Lacoss (1978) showed that the anti-alias filter originally installed on short period SRO instruments caused arrivals to be delayed almost 0.3 seconds relative to WWSSN stations. This bias is at least three times the current timing error for these instruments. So, in order to combine data from different digital acquisition systems, their group delays must be either comparable or accountable.

In order to use the existing CEDAR data, the third alternative given above; i.e., low pass filtering was tried as a temporary measure even though once the data has been digitized there is no way to determine or remove the actual effect of aliasing. It was hoped that low-pass filtering the seismograms would reduce the effects of aliasing when the signal spectrum extends only slightly above the Nyquist frequency.

The noise spectra of the short period instruments are generally fairly flat from 0 Hz. to the Nyquist frequency. Some instruments have prominent spectral peaks at 10 and/or 20 Hz. which is apparently due to aliasing of "sixty cycle hum" in CEDAR'S digitization stage. On most signals, the noise power was not significantly reduced unless a fairly narrow (0 to 10 Hz.) pass band was used. The effect of such narrow band filtering on most arrivals was usually quite small. However, events with magnitude below about two have appreciable amounts of energy above 10 Hz (Eaton, 1978). For such events, only about ten percent of the total energy is below 10 Hz. It was thus concluded that no prefiltering should be done on the CEDAR records as part of the algorithm for picking arrivals.

2.1. CEDAR OPERATION

Seismic data from the CEDAR network is processed in two steps. First the detection subsystem monitors the seismic network, detects events and records their seismograms on magnetic tape. Secondly, seismogists use a non-real time interactive subsystem to analyze the seismic data and produce a catalog of seismograms and earthquake locations.

Briefly, the detection subsystem works as follows. The network of 120 stations is divided into twenty overlapping subnetworks based on the known seismicity of the region and the network geometry. For each station, a running short term and long term average of the absolute value of its signal is determined. Whenever the ratio of the short term to long term average exceeds a certain factor the station is considered to have "triggered" indicating the possible presence of a seismic signal. A station remains triggered until the ratio returns below the factor. If five or more stations in any one subnetwork are concurrently triggered, an event is assumed to have occurred and seismic data from the entire network is saved.

Each day the seismic data collected by CEDAR for the previous day is analyzed by seismologists. The seismologists pick both P and S wave arrivals and locate the earthquake using an interactive program. The hand processing of CEDAR data will be discussed further in Chapter 3.

2.2. CEDAR STATION CHARACTERISTICS

Information provided by CEDAR was used to form a simple summary of the operating characteristics of each station as shown in Exhibit 2.5. At the moment that enough stations have triggered to cause CEDAR to record, the state of the detection parameters is saved. For each station the following information is saved:

1. DC bias

2. Short term average absolute value of signal

3. Long term average of absolute value of signal amplitude

4 Whether the station had triggered or not

The long term average essentially indicates the background noise level. The noise level at most stations remains fairly constant and is usually less than thirteen digital units. Each station has its own DC bias which varies slowly with time. The number of times a station triggers the network is an indication of the local seismicity around the station and the stations "sensitivity".

Based on Exhibit 2.5, the following stations were never used in automatic processing because they had either extremely large noise levels or very large deviations in DC level: coa, cok, crr, hdg, ing, run, scy, slu, sup. Also, four stations were never allowed to trigger the array, they are adl, blu, coq, and lhu. These stations were also not picked automatically.

2.3. EXHIBITS

2.1. The stations of the Southern California Seismographic network which are monitored by CEDAR.

2.2. Epicenters of the 63 events recorded by CEDAR from April 26, 1977 to June 2, 1977. The roughly 2500 seismograms, 1300 hand picked P wave arrival times, and 600 hand picked S wave arrival times from these events form the data base studied in this thesis.

2.3. Examples of aliased digital data recorded by CEDAR. The tick marks are 0.2 seconds apart and the sampling rate is 50 Hz. Aliasing is indicated by the presence of oscillations which have only one point per half cycle.

2.4. The Benioff stations monitored by CEDAR

2.5. Summary of the operating characteristis of CEDAR stations. Stations were not used in the automatic processing described in this thesis if they were very noisy, had large variations of their dc level, or were not used to trigger CEDAR.

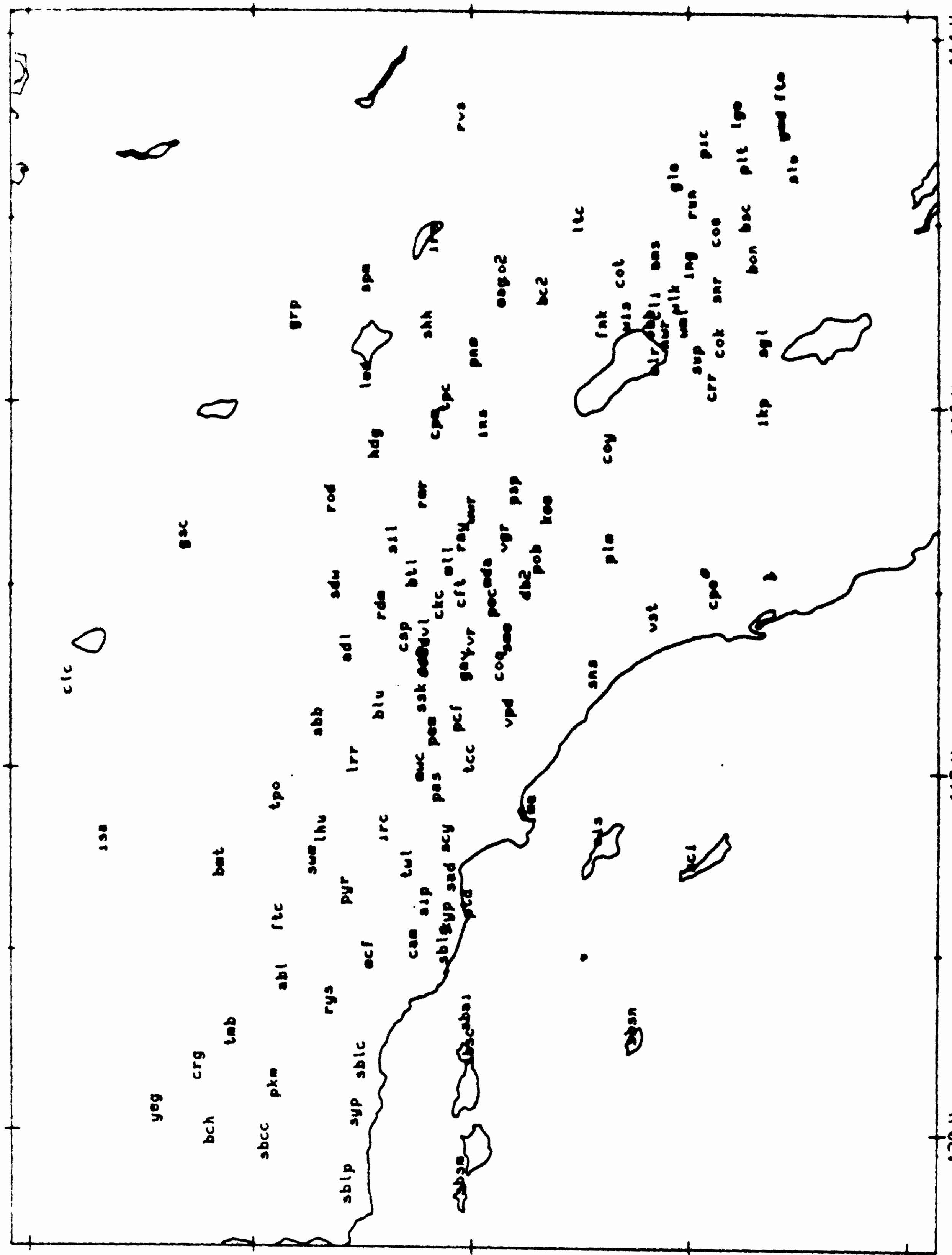


EXHIBIT 2.1

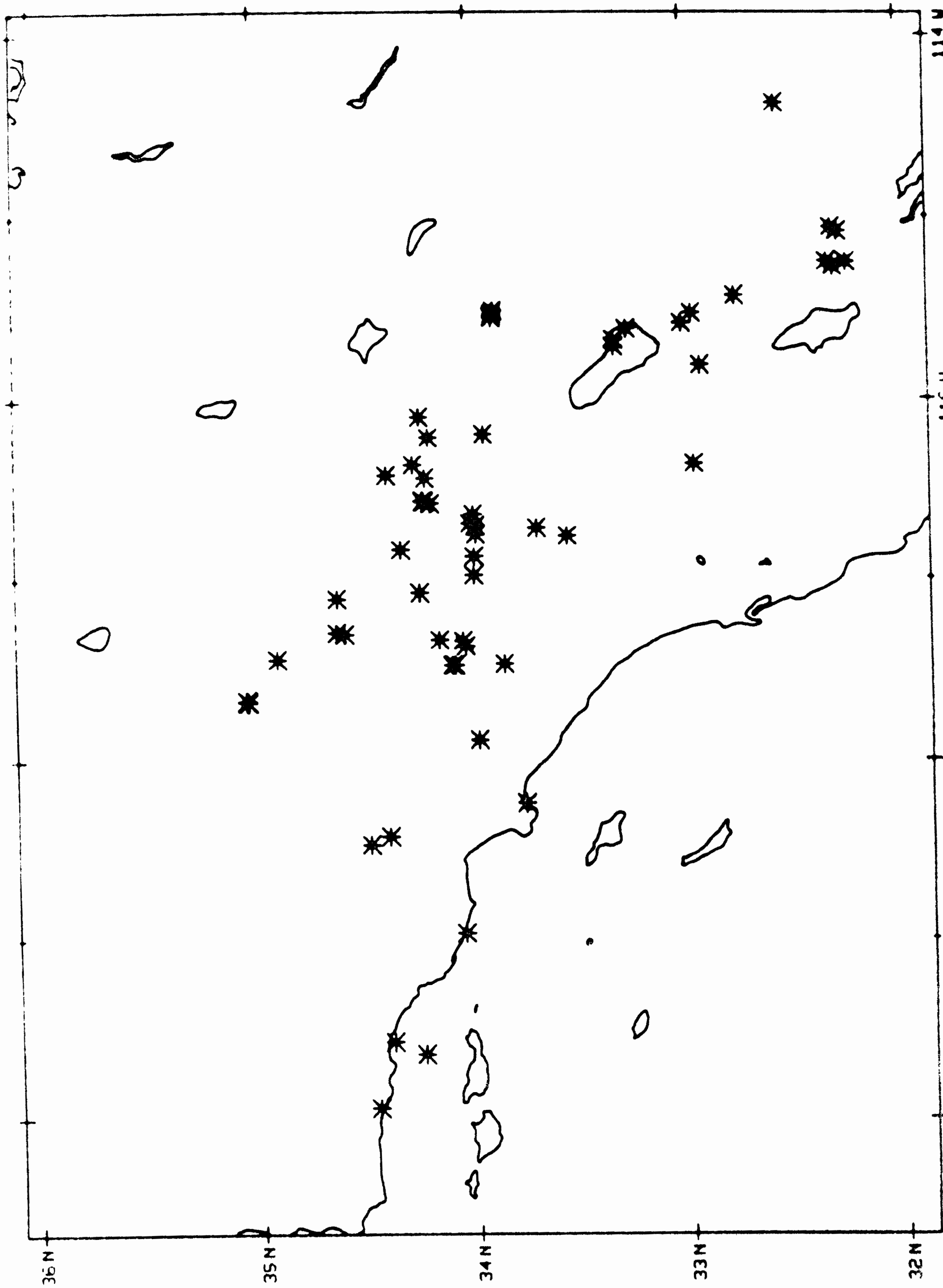


EXHIBIT 2.2

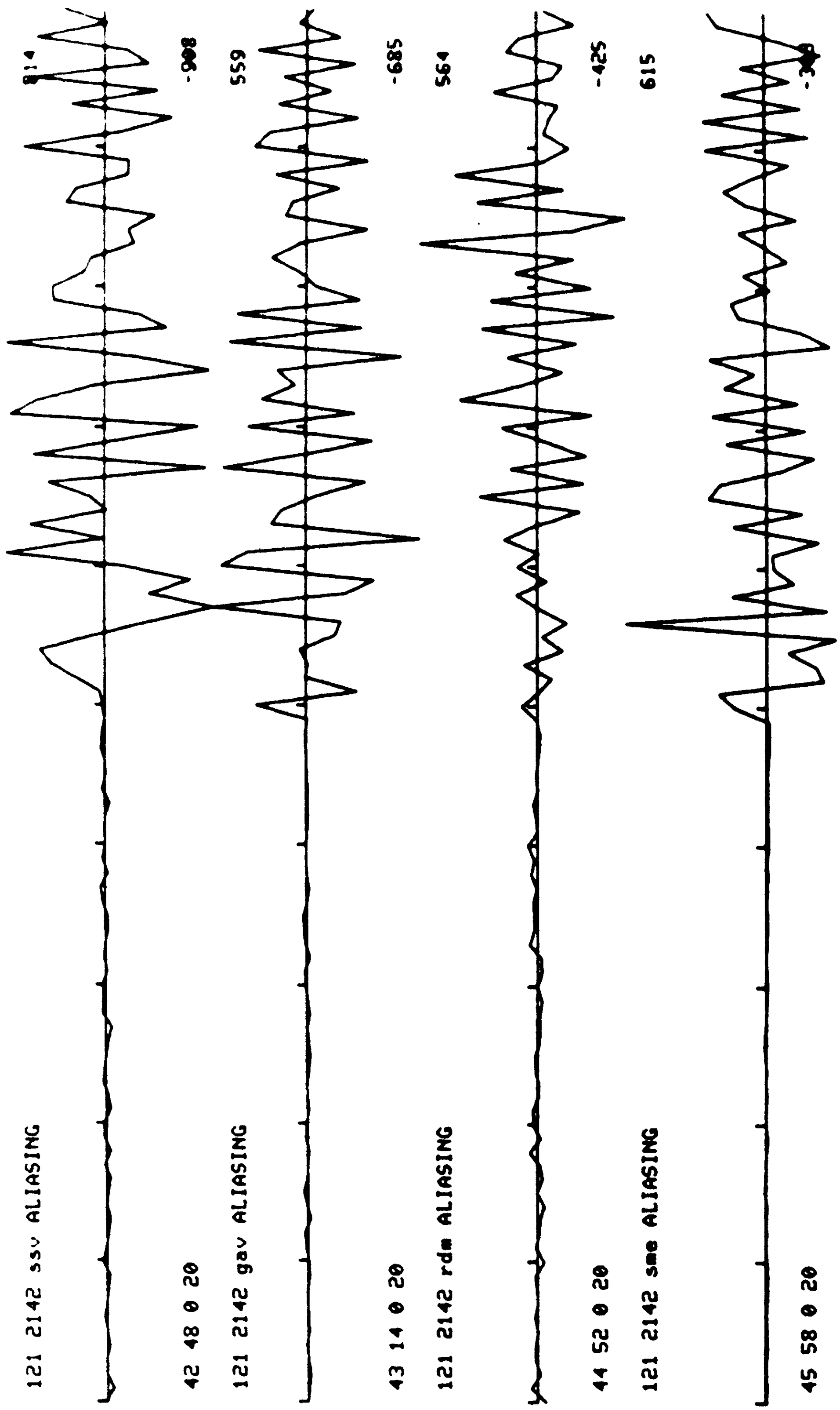

121 2142 ssv ALIASING
42 48 0 20
814
-908
121 2142 gav ALIASING
43 14 0 20
559
-685
121 2142 rdm ALIASING
44 52 0 20
564
-425
121 2142 sme ALIASING
45 58 0 20
615

EXHIBIT 2.3

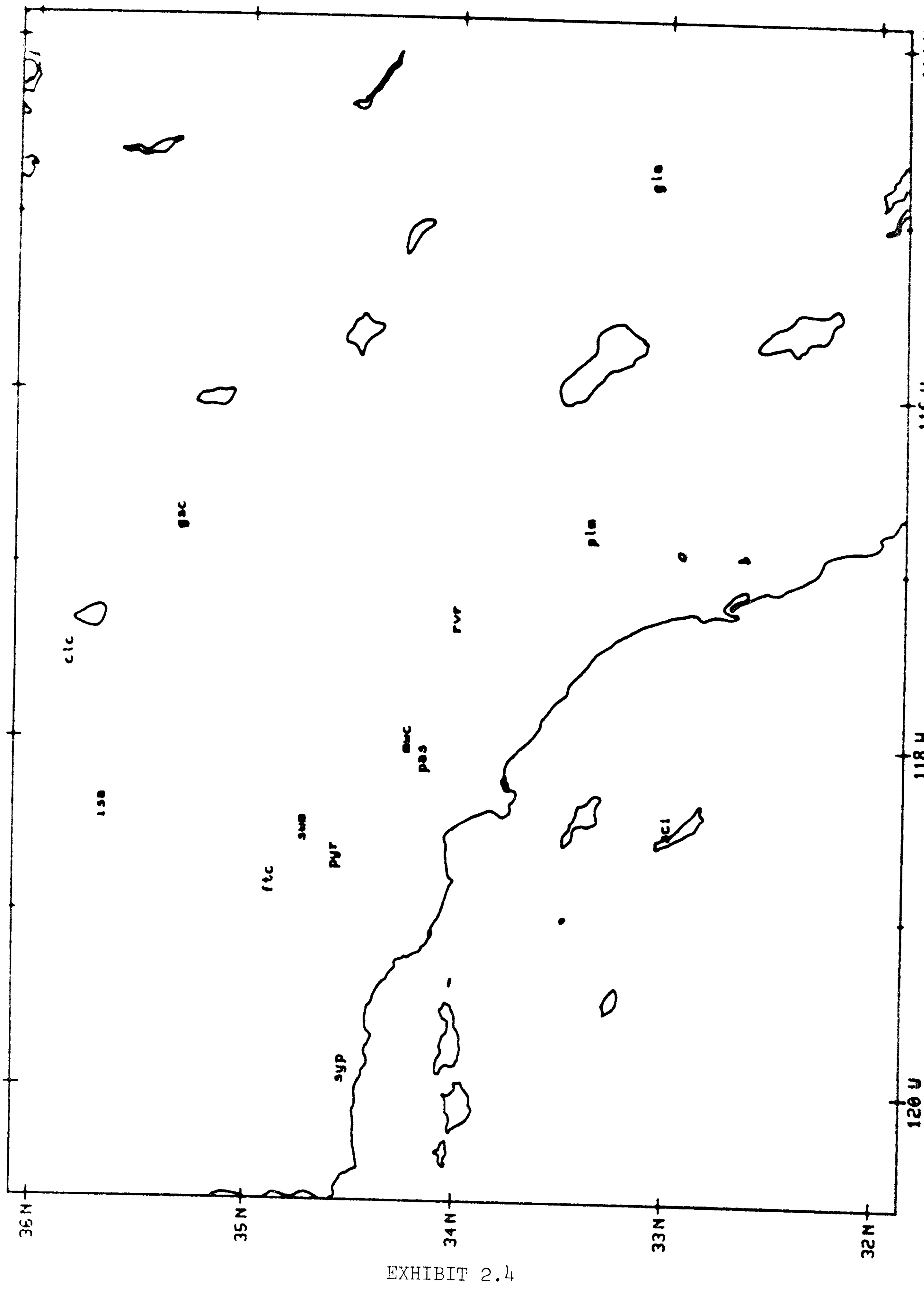


EXHIBIT 2.4

EXHIBIT 2.5

CEDAR STATION CHARCTERISTICS

columns are:
1 station code,
2 number of seismograms,
3, 4 dcbias median and scale
5, 6 long term average median and scale
7 number of times this station helped "trigger" the array
8 comment

where scale is the median(abs. difference from median)

1	2	3	4	5	6	7	8
abl	17	56	10	10	1	4	
adl	14	1	3	21	1	0	trigger off
ams	16	-222	97	17	2	2	large dc variation
bc2	30	-62	16	5	1	5	
bch	5	-72	9	9	1	0	
blu	31	-13	15	17	1	0	trigger off
bmt	19	171	11	6	1	4	
bon	10	-362	2	22	1	4	
bsc	10	-231	7	19	2	0	
btl	42	60	56	8	2	22	
cam	12	-234	8	19	4	1	
cft	27	15	40	7	1	4	
cis	23	118	3	9	2	1	
ckc	8	-147	13	6	0	0	
clc	18	0	1	3	1	4	
cli	11	-30	28	6	1	2	
co2	39	132	22	5	1	6	
coa	3	293	1	369	15	0	very noisy
cok	7	221	13	36	4	1	very noisy
coq	2	-229	19	1	1	0	trigger off
cot	22	45	1	9	2	4	
coy	1	22	0	4	0	1	
cpe	11	57	1	2	0	0	
cpm	35	231	23	8	1	8	
crg	3	-320	1	11	0	0	
crr	3	267	15	476	33	0	very noisy
csp	38	-156	13	6	0	9	
db2	50	42	20	5	1	21	
dc5	1	-122	0	0	0	0	
dvl	1	-49	0	1	0	0	
eag	7	-184	26	16	2	3	
ecf	21	-9	9	17	4	3	
elr	15	-60	20	10	2	6	
fma	8	-119	0	9	4	2	
fnk	12	30	50	7	1	5	
ftc	12	-199	29	6	0	0	
ftm	18	-101	26	5	1	7	
gav	39	112	28	9	2	10	

gla	29	11	1	6	1	2	
grp	31	-307	16	11	2	6	
gsc	27	-28	1	3	1	10	
hdg	24	-45	55	50	14	3	very noisy
ikp	32	11	3	4	1	0	
ing	13	444	240	18	6	2	large dc variation
ins	40	-13	13	16	1	12	
irc	32	21	3	4	1	2	
irn	18	105	31	8	1	1	
isa	13	18	0	3	0	5	
kee	16	140	11	11	1	2	
kyp	2	-168	59	2	1	0	
lcl	8	55	5	14	1	1	
led	6	-252	0	19	5	1	
lga	14	-6	7	8	2	4	
lhu	20	16	15	11	1	0	trigger off
lrr	22	20	20	23	1	4	
ltc	30	-31	24	5	0	2	
mda	36	92	22	8	2	10	
mll	21	78	24	24	2	12	
mwc	27	-72	3	3	1	1	
nwr	6	-15	0	5	0	5	
obb	8	-56	1	18	2	2	
pas	18	-7	1	2	1	0	
pcf	10	-99	51	10	1	0	
pec	47	-4	3	5	1	10	
pem	42	-3	10	4	1	8	
pic	21	-21	8	5	0	9	
pkm	13	240	23	5	1	2	
plm	44	-111	1	4	0	2	
plt	19	139	9	15	3	7	
pnm	36	-292	20	4	1	6	
pob	29	19	9	23	6	4	
psp	25	-111	15	9	2	8	
ptd	20	-121	30	17	4	2	
pyr	26	-7	1	5	1	2	
ray	42	48	10	6	0	20	
rdm	42	-39	7	8	2	18	
rmr	33	-207	39	15	2	15	
rod	38	-21	30	10	0	14	
run	20	-68	96	14	2	6	large dc variation
rvr	29	-6	11	3	1	4	
rvs	27	-152	54	10	2	2	
rys	16	-158	18	7	0	4	
sad	21	-94	40	8	1	3	
sbai	8	-61	7	10	2	2	
sbb	31	220	13	10	2	9	
sbcc	9	-124	22	4	1	0	
sblc	13	37	28	3	0	3	
sblg	16	213	36	14	1	1	
sblp	5	-244	4	6	1	0	
sbsc	7	-78	3	20	1	0	
sbsm	7	-107	6	16	8	1	
sbsn	1	-33	0	7	0	0	

sci	5	11	3	6	2	0	
scy	2	890	222	6	2	0	large dc variation
sdw	23	-26	25	12	0	0	
sgl	29	-19	12	13	3	13	
shh	28	-170	18	13	1	3	
sil	34	161	41	11	3	17	
sip	27	-13	25	13	2	4	
slu	27	-22	6	44	12	4	very noisy
sme	45	-139	9	3	1	9	
snr	5	-34	0	11	0	4	
sns	30	13	2	8	2	0	
spm	33	-60	13	6	1	1	
ssk	1	-222	0	1	0	0	
ssv	40	-75	31	4	1	12	
sup	29	187	4	19	11	12	noisy
swm	30	-162	19	9	1	5	
syp	12	-473	4	5	1	2	
tcc	22	116	5	15	2	2	
tmb	6	-95	3	8	1	0	
tpc	36	-16	2	3	0	3	
tpo	27	3	33	7	1	8	
twl	14	45	4	9	2	3	
vgr	24	33	14	15	1	6	
vpd	29	-21	1	10	5	4	
vst	33	-1	8	5	2	0	
wis	6	66	30	15	5	3	
wlk	13	293	21	19	4	1	
wml	16	26	10	19	9	7	
wwr	25	37	26	6	0	8	
wwv	70	248	7	361	4	0	trigger off
yeg	10	124	17	9	0	0	
ymd	20	-28	2	5	1	8	

3. HUMAN DETERMINATION OF ARRIVAL TIME

This chapter is the first of two chapters concerned with the problem of determining the arrival time of seismic waves. Before a machine can be programmed to pick arrivals, one must first understand how seismologists pick arrivals. Picking arrivals is an art which is usually handed down from seismologist to seismologist by an informal period of apprenticeship. Unfortunately, beyond this apprenticeship there is very little discussion or comparison of how or why arrivals are picked. Only a few publications have been concerned with teaching how to interpret seismograms (Simon, 1972; Anon., 1966; Neumann, 1966; Richter, 1958, p. 290-295). Most of these publications consist of seismograms of earthquakes from different distance, depths and source regions so that the differences and similarities between the waveforms can by studied. Unfortunately, all of these publications are primarily concerned with teleseismic arrivals and give very few examples of local seismograms like those recorded by CEDAR. There is little discussion on how to identify the actual onset of an arrival or how to assign an accuracy estimate to it.

Since picking arrival times is a heuristic [1] process, each person is bound to have his own set of biases. These biases have been considered in some detail by Freedman and her results will be summarized in the next section. The section following will compare hand picked arrivals from CEDAR with Freedman's work.

---

[1] The term "heuristic" means a rule of thumb or a form of advice. Heuristics are used whenever there is no complete algorithm or if the algorithm is too expensive to use. People use heuristics to reduce the complexity of making decisions under uncertainty. Reliance on heuristics, however, can lead to systematic biases which must be taken into account (Tversky and Kahneman, 1974).

3.1. THE PSYCHOLOGY OF MEASURING ARRIVAL TIME

As part of the development of the Herrin travel time tables, Freedman studied the accuracy of picking seismic arrivals (Freedman 1966a, 1966b, 1968). Although she used teleseismic data in her experiments, her results are probably general enough to apply to almost any seismological timing problem.

Freedman used nine experimental subjects with a wide range of experience in reading seismograms. Each subject was to determine the onset times of all arrivals occurring on six days of seismic data and to describe the arrival as "impulsive" (sharp onset) or "emergent" (gradual onset). Each subject read the dataset a total of three times with usually a month between each reading. Since real seismic data was used, it was impossible to determine which picks actually corresponded to a true arrival. It was assumed that if an arrival was reported at least three times by at least two different readers then it was a genuine arrival. The median arrival time reported for an arrival was used as the true arrival time. All other arrivals were considered false alarms.

The experiment showed that each reader has their own set of individual biases in what they consider an event to be and what they pick as the onset time. Experienced readers had a different set of biases than novice readers. More experienced readers appeared bolder. They tended to pick a larger number of arrivals which no other reader saw. The novice readers were more cautious. They tended to pick only the most unambiguous arrivals and miss weaker arrivals. And since they pick mostly impulsive arrivals, the standard error in their arrival times is usually smaller than for the bolder readers.

A simple explanation of this can be given by analogy to a binary decision problem. The cautious reader uses a higher confidence threshold [1] and therefore will not pick the weaker arrivals. With experience, the reader gains additional confidence and hopefully skill. Their personal threshold is lower. They now pick weaker arrivals but also pick more false arrivals. This tradeoff is a well known aspect of statistical decision theory (Helstrom,1960).

Although the criterion used by readers for discriminating signals from noise varies considerably from one reader to another it is used by the readers themselves with a fair degree of consistency. Of the 60 percent of arrivals seen by only one reader, only 40 percent were seen on only one of the three readings.

The description assigned to the arrivals also showed the readers personal bias. In fact, even though two subjects read almost exactly the same events the proportion of events that they identified as impulsive was 0.18 and 0.68 respectively.

Six of the nine readers were biased to pick arrivals later than the median arrival time. This is a significant problem since the arrival times are most important in determining the location of earthquakes and the structure of the earth. Freedman (1966a, 1966b, 1968) suggested that the effect of personal bias can be reduced by having two people read each seismogram and use only those arrivals for which both agree.

---

[1] The term "threshold" is used only to make the analogy to the standard binary decision problem. The actual method used by people to discriminate between noise and seismic signal is obviously quite complex and cannot be explained by only a single parameter or threshold.

However, this does not guarantee that gross errors such as overlooking the first arrival will be completely eliminated. While studying travel times from Nevada Test Site blasts recorded in California, Freedman (1966b) observed that even when several readers obtained estimates differing by no more than 0.1 seconds, the travel time residuals based on them were often larger than one second. Both readers had picked the same wrong arrival. On the average, the more energetic explosions in tuff or granite tended to produce earlier readings and smaller residuals. The same effect has been observed by Steppe, Bakun and Bufe (1977) and Lindh, Lockner, and Lee (1978). They showed that at some stations, smaller events had later residuals than larger events. In fact, for smaller events (less than magnitude 1.8) arrivals can be as much as 0.6 seconds later than arrivals for the larger events.

Pearce and Barley (1977) have shown qualitatively how noise affects the appearance of the first half cycle of a seismogram. By adding different levels of noise to synthetic seismograms they showed that the first motion can be completely hidden or appear to have the opposite sense of motion. They feel that the apparent reliability of measurements made on signals in noise is often an illusion and suggest that a signal to noise ratio (S/N) of greater than 6.0 is required before the first motion should be classified reliabily.

No quantitative study of how accurately people pick arrivals has been made for the microearthquake data studied in this thesis. However, Lee (W. H. K. Lee, personal communication, 1978) has kept informal records of arrival times picked by U.S.G.S. personnel during training. For impulsive arrivals he found that people usually agree to within 0.05

to 0.1 seconds in arrival time and to within 25 percent on the quality of the arrival.

In summary, Freedman has shown that there is a significant amount of personal bias in picking onset times and in describing first arrival waveforms. Although her methods remove most of the effects of this bias it is still possible that the wrong arrival is chosen. This is supported by Pearce and Barley's work who show that the quality of an arrival is sometimes an illusion. This is a significant problem since averages of many arrivals are used to determine structure, station corrections and VP/VS ratios and so on. Freedman (1966b, p. 681) has pointed out that any average which is contaminated by arrival times from two different phases is useless. The fact that weak arrivals tend to be read late is also a problem since the current theories of premonitory velocity changes predict only lower velocity (Lindh, Lockner, and Lee, 1978). Lindh, Lockner and Lee (1978) have shown that at least one recent report of velocity decrease before an earthquake was based primarily on weak ambiguous arrivals.

3.2. CEDAR HAND PICKS

As described in Section 2.1 , only the data acquisition stage of CEDAR is currently automated. Arrival times are determined manually using an interactive computer program which displays and manipulates the data for the seismologists. Picking arrivals interactively has been shown to be faster and more accurate than previous methods using data recorded on microfilm.

Arrival times on microfilm are sometimes difficult to read since data is displayed at a fixed magnification, and traces tend to overlap and obscure each other. Also, since the data is recorded on the film by a light beam, rapid motion may be underexposed. Arrival times must be measured from the microfilm using a ruler and transfered to punched cards manually. Thus errors due to parallax and key punching errors are likely.

With the interactive program, there are no parallax or mispunching errors since a computer graphics display is used, and the computer takes care of all the bookkeeping. Since the seismologist can control the scale that the data is displayed, arrivals are more accurate. In fact, it seems that descriptive terms like "emergent" must now be redefined (Carl Johnson, personal communication 1977).

Each arrival is given a description from a standard set of descriptions (Lee and Lahr, 1975) . The arrival is described as implusive (I) if the onset is sharp; or emergent (E) if the onset is gradual. The direction of first motion is either up (U) for compression or down (D) for dilatation (If the analyst is uncertain in the direction of first motion the symbols "+" or "-" respectively are used.). Each arrival time is given a quality weight from 0.0 to 1.0 which indicates how reliable the arrival time is. This weight is used in the earthquake location scheme as described in Chapter 5. In this thesis the terms Q0, Q25, Q50, Q75, and Q100 will be used to represent quality weights of 0.0, 0.25, 0.50, 0.75, and 1.0.

The CEDAR database studied in this thesis contains nearly 1300 hand picked P-wave arrivals and 600 S-wave arrivals. These arrivals will be used as a baseline to which the machine picked arrivals will be compared. Although the arrivals were picked by one or two seismologists, it will be assumed that they are representative of the arrivals picked manually for the CEDAR network.

The distribution of hand picked arrival descriptions for both P and S waves are shown in Exhibit 3.1. Roughly half of the P wave arrivals are assigned qualities of Q100 and Q75 and half are assigned qualities of Q50 and Q25. Ignoring the Q0 arrivals which are not used in locating the earthquake, all but one of the Q100 and Q75 arrivals are described as impulsive and all of the Q50 and Q25 arrivals are described as emergent.

The confidence in correctly identifying the sense of first motion decreases with the quality. The operators are confident in all of the Q100 arrivals, and only half of the Q75 arrivals; they are uncertain about the first motions for the Q50 and Q25 arrivals. In fact, most of the lower quality arrivals have no first motion specified.

Thus the quality weighting provides most of the information about the arrival: Q100 and Q75 arrivals are impulsive and their first motions are usually reliable; Q50 and Q25 arrivals are emergent and their first motion is relatively unreliable or unspecified.

Generally, S arrivals are more difficult to identify than P arrivals since they occur in the coda of the P arrival, and thus they are given lower quality estimates as shown in Exhibit 3.1. There are very few Q100 arrivals and the rest are about evenly distributed between Q75, Q50, and Q25. Also, most S arrivals are considered emergent and do not have first motions recorded for them.

A histogram of the travel time residuals [1] for hand picked p wave arrivals for each quality category is shown in Exhibit 3.2. The histograms have been overlaid by curves which are the sum of two Gaussians to approximate the shape of the underlying probability distribution (see Exhibit 3.3). The Q100 arrivals are almost pure Gaussian with a σ of 0.1 sec. Arrivals of the lower quality seem to be fit fairly well by Gaussian distribution contaminated by a Gaussian with a larger standard deviation. For the lower quality (Q50 and Q25) arrivals at least, larger residuals tend to be late arrivals. These may be due to weak arrivals in which the first half cycle was not picked. For all qualities, a few percent of the residuals are quite large.

A scatter plot of residuals vs distance is shown in Exhibit 3.4. There is a fairly dense cloud of small residuals from 0 to about 100 km. distance superimposed on a fairly uniform background of larger

[1] The residuals are from the BI location method described in Section 5.3

residuals. The relative travel times for the Kanamori and Hadley (1975) velocity model for Southern California are also shown.

A somewhat more informative view of this data is shown in Exhibit 3.5. This type of exhibit will be referred to as a "median-hinge summary" and was formed as follows: The arrivals were sorted by distance and divided into groups of forty nine. Each group was then sorted by residual. The lines show the median and upper and lower hinges (25 percent and 75 percent quartile points) plotted at the median distance of each group (Tukey, 1977).

This exhibit shows that from 0 to 120 km. absolute value of most of the residuals are less than 0.2 seconds and the median residual is near zero. From 100 to 160 km. the median residual is 0.16 seconds late and the upper hinge curve rises sharply, indicating that this region is dominated by late picks. At this distance, the first arrival is usually fairly weak, and the actual onset may be missed (Roller and Healy, 1963). This increases the chance of the mantle reflection (the dashed line in Exhibit 3.4) being misinterpreted as the first arrival. Borcherdt and Healy (1968) have also found that the variance of residuals increases with distance.

Beyond 160 km. the median and both hinges drop rapidly. At this distance range the first arrivals are refracted through the mantle and have a velocity of about 7.8 km./sec. (Kanamori and Hadley, 1975). Thus beyond 160 km. the constant 6.25 km./sec. velocity medium used to locate the events is an inadequate velocity model and a layered model should be used. However, since only about ten percent of the arrivals occur at distances greater than 160 km. and the constant velocity model is quite adequate for most events.

To summarize how seismologists pick arrival times: Q100 and Q75 arrivals are impulsive, their first motion estimate is usually considered reliable, and arrival times usually agree to within 0.05 to 0.1 seconds from seismologist to seismologist. Q50 and Q25 arrivals are emergent and there is some tendency for the arrivals to be late. For two thirds of these arrivals, no first motion is given indicating that the time of onset is ambiguous by a half cycle or more. Beyond 100 km. the spread of the residuals increases and there is a definite bias toward late arrival times because the first arrival is weak and may be confused with later arrivals.

3.3. EXHIBITS

3.1. The distribution of hand picked P and S wave descriptions (See text).

3.2. Histograms of the travel time residuals (arrival time minus theoretical arrival time) for hand picked arrivals of each quality type. For comparison, the dashed curves are the sum of two Gaussians. The parameters of the Gaussians are shown in Exhibit 3.3.

3.3. The approximate distribution of hand picked residuals for each arrival quality. The notation $N(\mu,\sigma)$ stands for a Gaussian with mean, μ, and standard deviation, σ.

3.4. A scatter diagram of travel time residuals for hand picked arrivals as a function of distance. The relative travel times for the Kanamori and Hadley (1975) velocity model are also shown. The solid lines are refracted arrivals and the dashed line is the mantle reflection.

3.5. A median-hinge summary of travel time residuals for hand picked arrivals. This exhibit was constructed as follows: The residuals were sorted by distance and divided into groups. Each group was then sorted by residual. The solid line is the median and the dashed lines are the upper and lower hinges (25 percent and 75 percent quartile points) of each group plotted at the median distance of each group (Tukey, 1977).

EXHIBIT 3.1

DISTRIBUTION OF HAND PICKED
P-WAVE DESCRIPTIONS

	QUALITY 100	75	50	25	0	MOTION
I	234	83			8	U
I	120	51			10	D
I		69			1	+
I		59			1	-
I	2	1			1	?
E						U
E						D
E		1	68	44	5	+
E			66	23	5	-
E			172	188	72	?
	28	20	23	20	8	% of total

DISTRIBUTION OF HAND PICKED
S-WAVE DESCRIPTIONS

	QUALITY 100	75	50	25	0	MOTION
I	5	29			2	U
I	4	28				D
I						+
I						-
I	10	82			2	?
E						U
E						D
E			8			+
E			22	2		-
E			150	156	24	?
	4	27	34	30	5	% of total

EXHIBIT 3.2-a

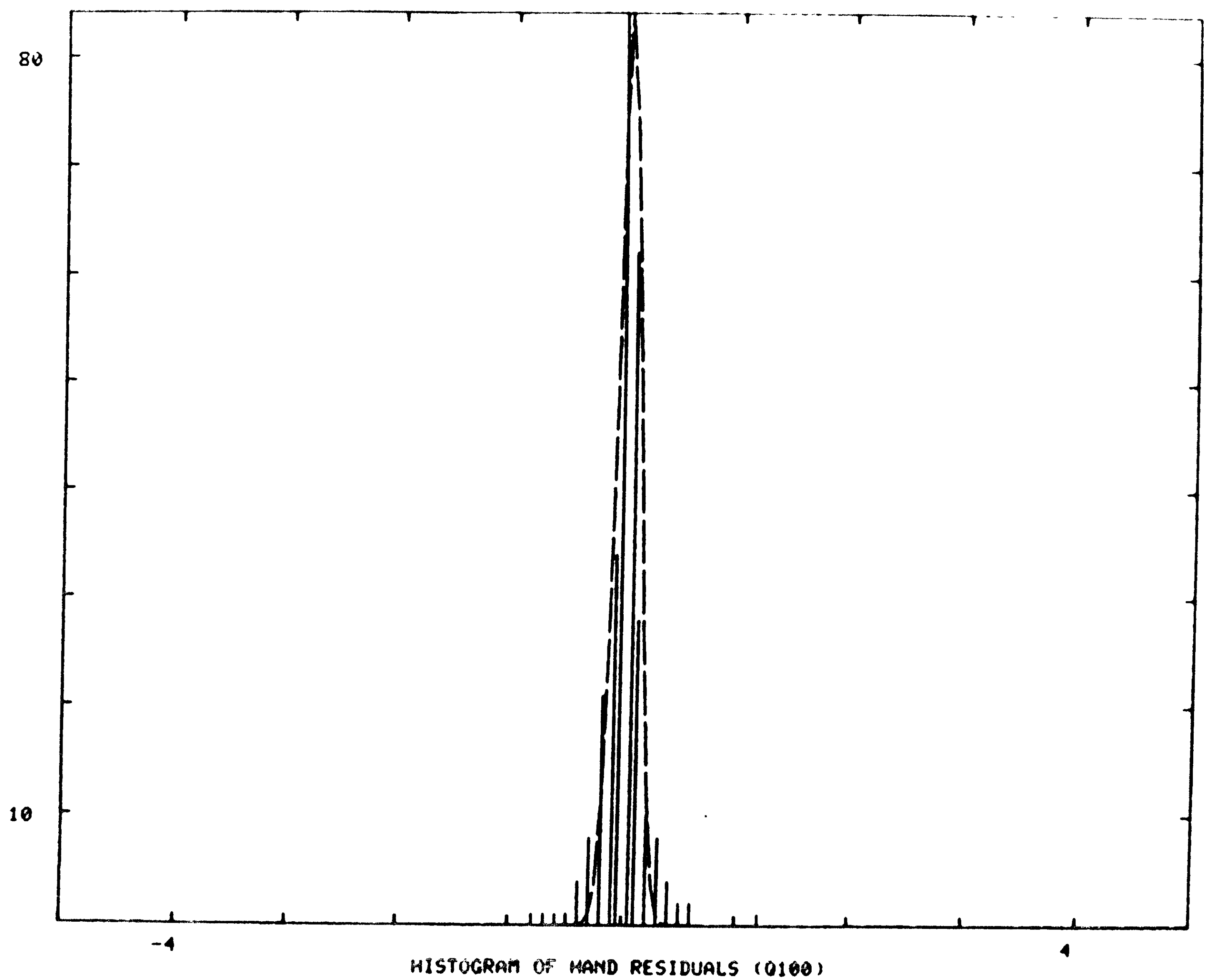
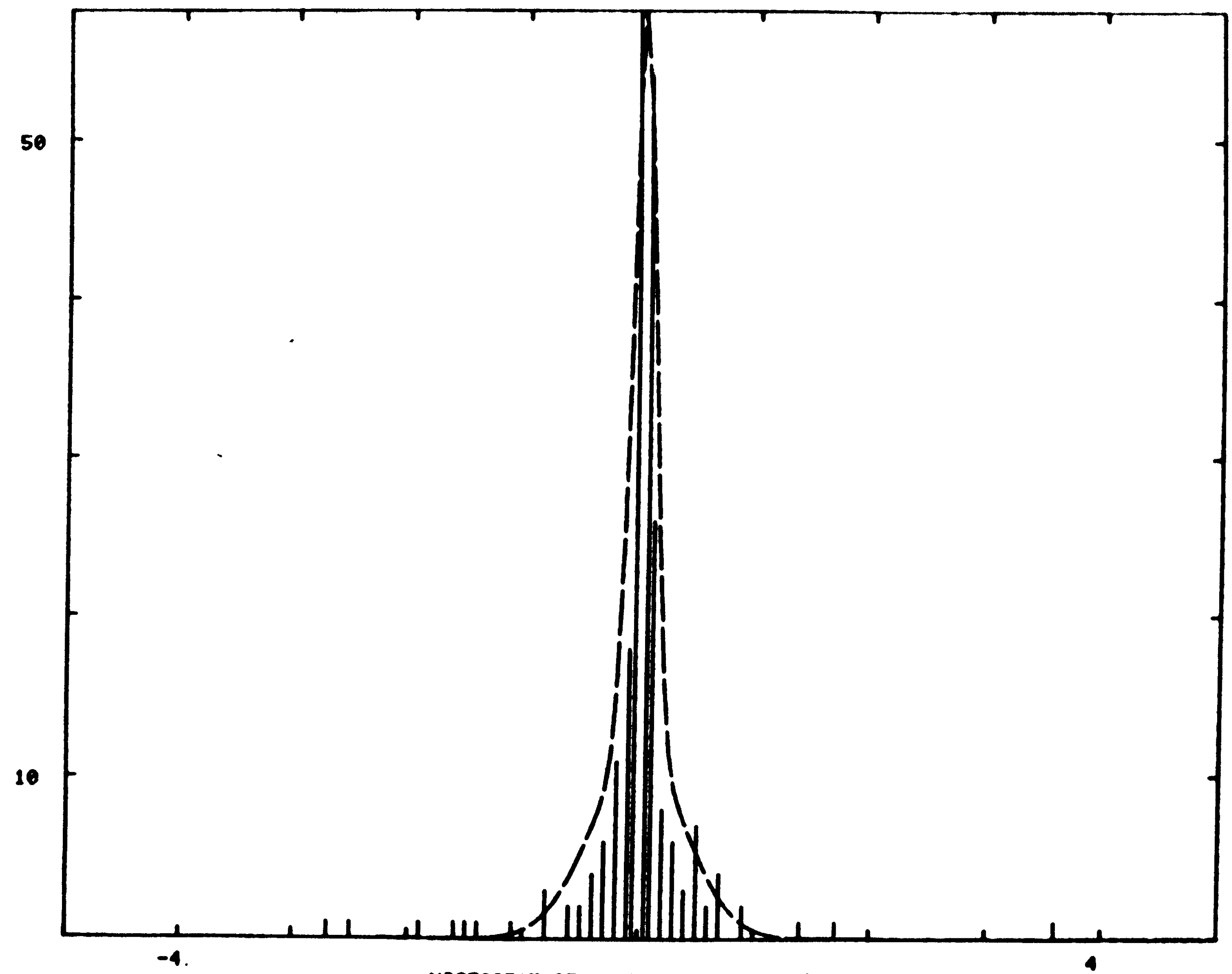
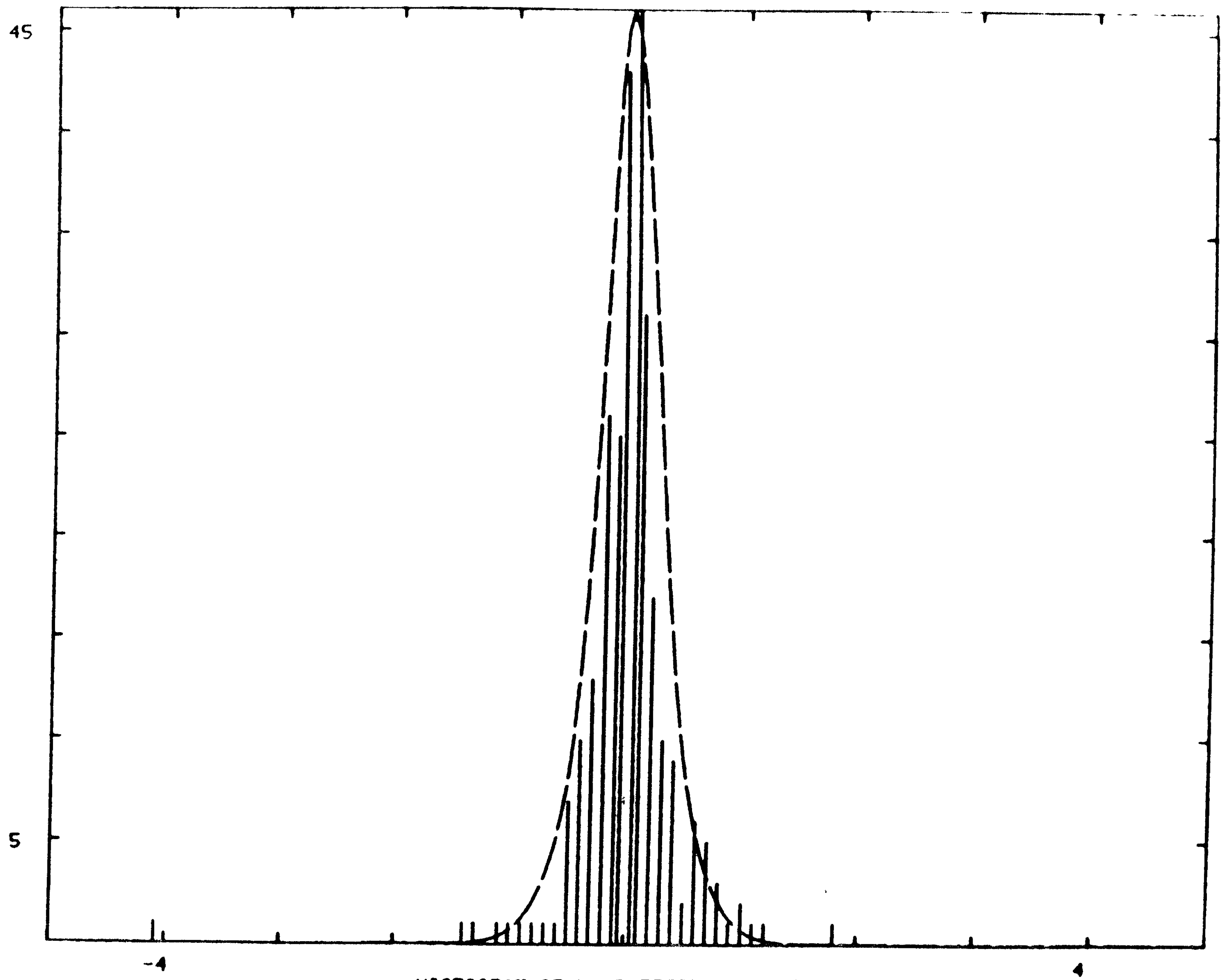


EXHIBIT 3.2-b



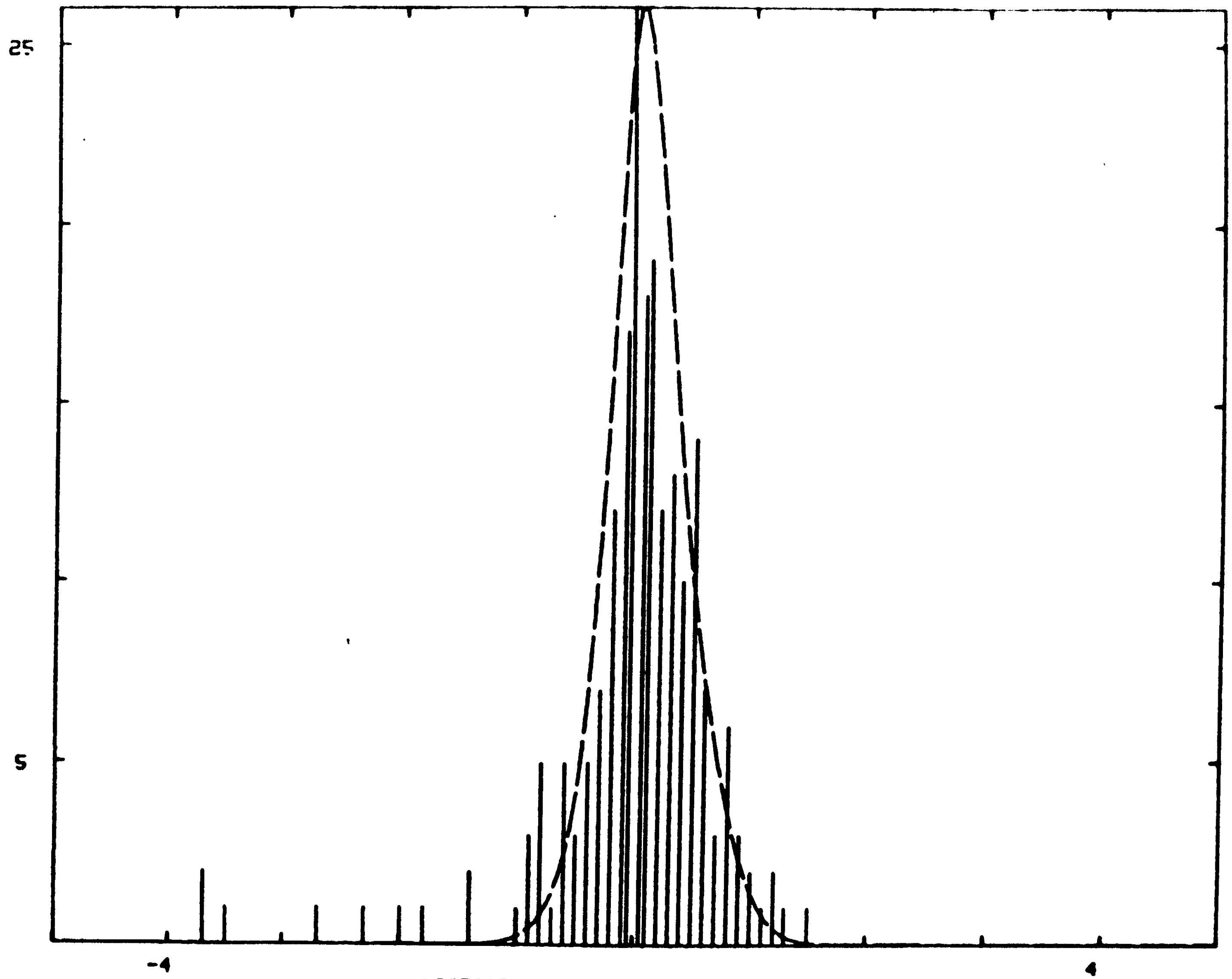
HISTOGRAM OF HAND RESIDUALS (Q75)

EXHIBIT 3.2-c



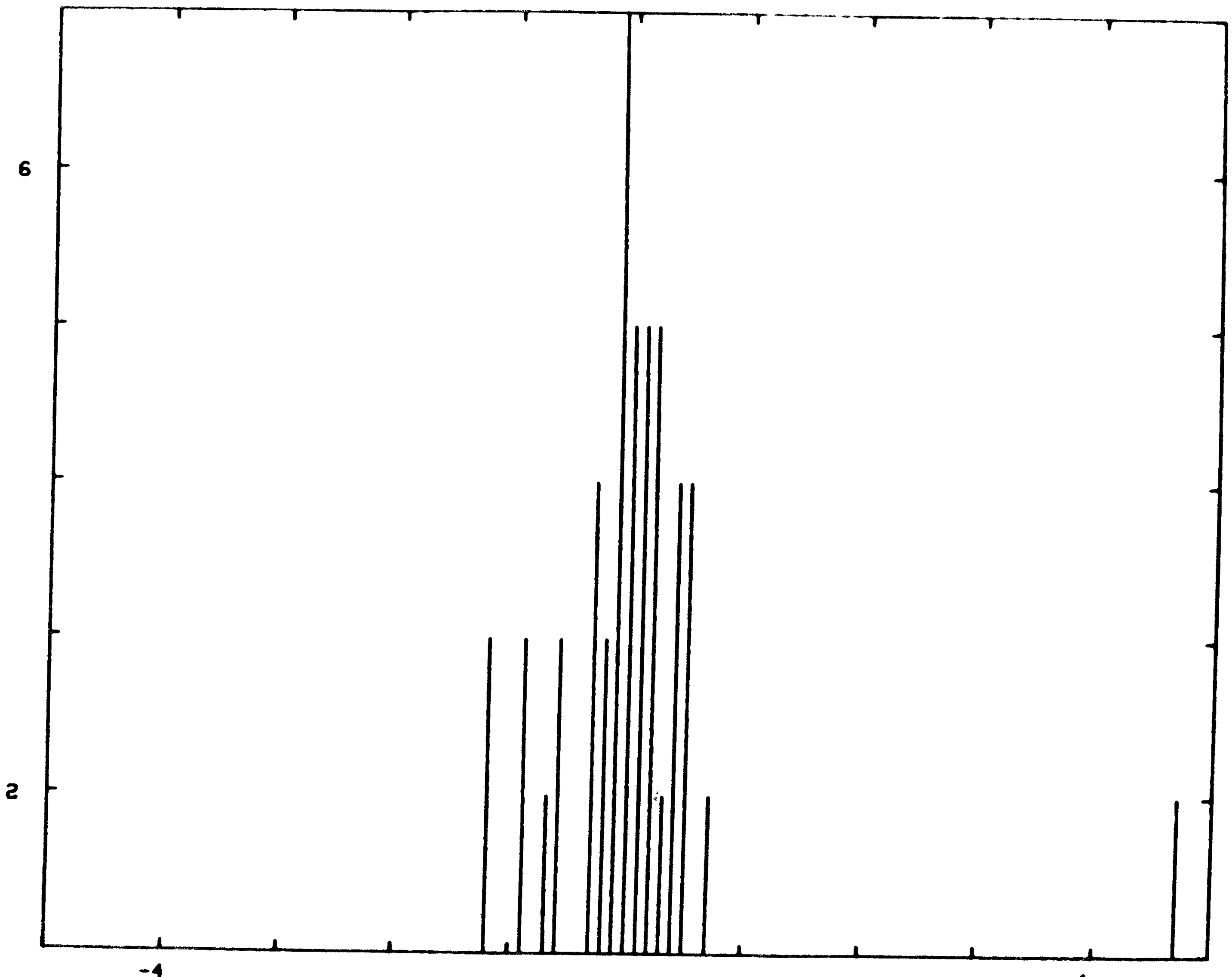
HISTOGRAM OF HAND RESIDUALS (Q50)

EXHIBIT 3.2-d



HISTOGRAM OF HAND RESIDUALS (Q25)

EXHIBIT 3.2-e



HISTOGRAM OF HAND RESIDUALS (00)

EXHIBIT 3.3
APPROXIMATE DISTRIBUTIONS OF HAND PICKED RESIDUALS

QUALITY	APPROX. DISTRIBUTION
100	N(0, .1)
75	0.8N(0, .1) + 0.2N(0, .4)
50	0.66N(0, .2) + 0.33N(0, .4)
25	0.5N(0, .2) + 0.5N(.15, .4)

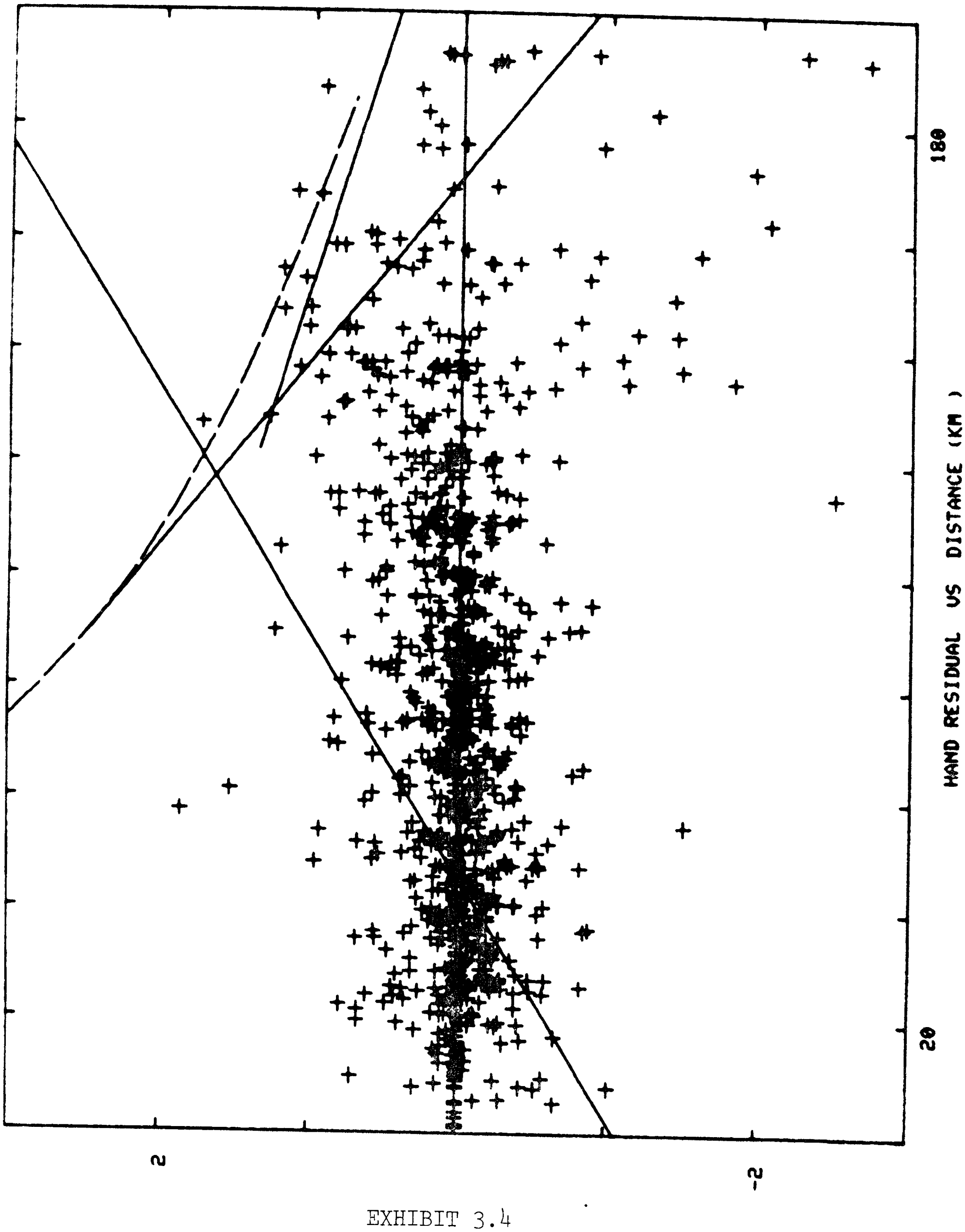
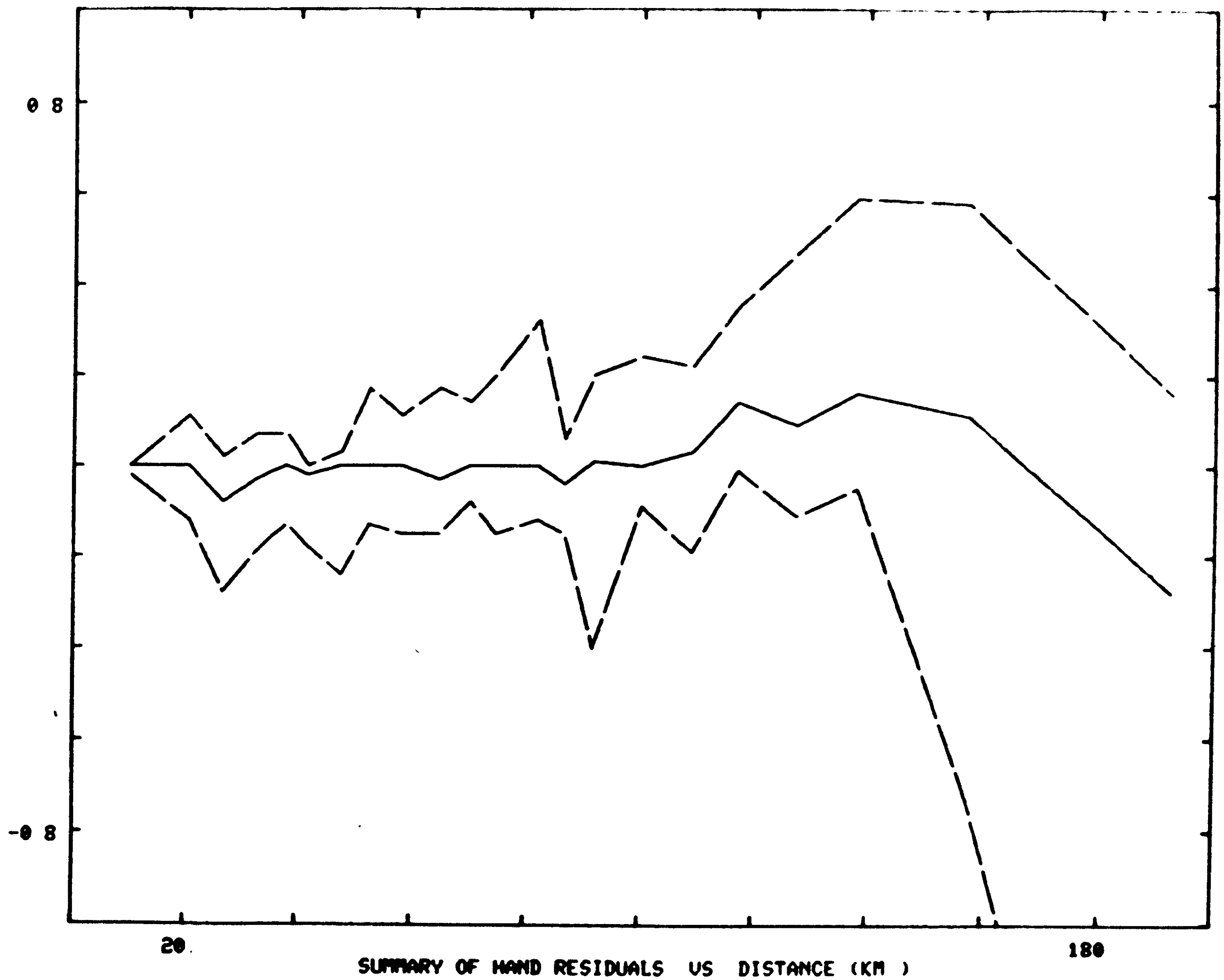


EXHIBIT 3.4

EXHIBIT 3.5



4. AUTOMATIC DETERMINATION OF ARRIVAL TIME

This chapter is concerned with automatically picking arrival times. Determining a signal's arrival time is important in a variety of scientific disciplines. It has been studied extensively in the fields of radar, sonar, and communication where a large amount of detection theory has been developed (see for example Woodward, 1957 p. 62-125; Helstrom, 1960 p. 203-330; Mityashev, 1965 p. 1-24, 123-182; and Van Trees, 1971). Unfortunately, in local seismology, the signal to be detected is sufficiently complicated that detection theory is only of qualitative value.

We begin by looking at the variety of signals with which any picking algorithm will have to deal. Exhibit 4.1 shows a suite of seismograms from event 116.414 (coda magnitude 2.2) which is typical of the medium sized events recorded by CEDAR. The seismograms range in distance from 8 to 150 km. and show the effect of distance on a seismogram's shape. The P and S wave codas are visually separate for distance beyond about 30 km. but their relative amounts of energy are quite variable and depend on the source mechanism and propagation path. Generally the amplitude and dominant frequency of the seismograms decrease with distance due to geometrical spreading and attenuation. First arrivals are usually impulsive and easy to identify near the event becoming progressively more emergent and ambiguous with distance (see Chapters 3 and 6).

Individual stations often have a strong influence on the shape of their signals (Reiter and Monfort, 1977). For example, station cft always has arrivals more emergent than any station near it even if cft is near the event. Also, some stations have noise characteristics which make identification of arrivals on them more difficult. Generally, first arrivals are variable from station to station even for stations which are near each other.

Unfortunately, the nature of seismic noise is at least as varied as that of the seismic signals we wish to record. Thus, any automatic process for recognizing seismic signals must spend a great deal of its effort handling seismic noise, (Stewart, 1977).

The most prevalent noise is the more or less Gaussian background noise. The noise spectra of the short period instruments are generally flat from 0 Hz. to the Nyquist frequency (25 Hz.). Some instruments have prominent spectral peaks at 10 and/or 20 Hz. which is due to aliasing of sixty cycle hum in CEDAR's digitization stage.

Besides the Gaussian background noise, there are at least four types of non-Gaussian noise which occur on the array; glitches, DC offsets, noise bursts, and cross talk. Glitches are short segments (generally one or two samples long) which are of significantly different amplitude than their surroudings, see Exhibit 4.2. CEDAR data has at least three different causes of glitches (Carl Johnson, Personal Communication 1978):

1. Telephone line or A-D converter glitches which occur on several channels simultaneously.

2. Certain channels characteristically have larger numbers of isolated glitches.

3. Several stations, such as BMT have long series of glitches-- one glitch every five samples. These glitch trains have slowly varying amplitudes and are due to aliasing of 60 cycle hum at the A-D conversion stage.

DC offsets are steps in the DC level as shown in Exhibit 4.3. The DC level jumps suddenly and then slowly returns to its previous level. Noise bursts come in various forms, a high frequency example is shown in the second trace of Exhibit 4.3.

Cross talk occurs when a small portion (typically five percent) of one station's signal appears on another channel. It is a particularly diabolical type of noise since it is correlated with the seismic event. Two examples of cross talk are shown in Exhibit 4.4. In each example, the first two traces are a blow-up of the first five seconds of the following two. Note that the first two traces in Exhibit 4.4-a are quite similar except that the second is twenty times larger than the first. The second trace is the S arrival on station sbb which is about 40 km. from the event. The first trace is the cross talk on station mwc from sbb. Since mwc is 100 km. from the event the weak seismic arrival is easily hidden by the cross talk. These seismograms are from a quarry blast which occurs daily at approximately local noon. For each blast, the mwc signal is contaminated by sbb and hand picked arrival times from mwc are typically three seconds or more early.

An example of where station btl talks to rdm is shown in Exhibit 4.4-b. In this case, both stations are near the source and thus the cross talk does not strongly affect the actual first arrival. Unfortunately, this arrival can be picked by an automatic algorithm as a weak first arrival.

It is relatively easy to detect the presence of seismic signals as their energy and spectral content is usually significanty different from the background noise and their duration quite long. Traditionally, comparing a short term average of the input signal power to a long term average of the signal power has been used as a simple detector of seismic energy. Such a detector is used in CEDAR's detection stage as described in (Section 2.1). It has proven successful for event recording seismometers (Ambuter and Solomon, 1974) and has been used as part of several automatic systems (Shlien, 1972; Allen, 1976; Stevenson, 1976). It has the advantage of being simple to implement in either hardware or software, and is adaptive to noise fluctuations.

At this point the distinction between "detection" and "identification" (Marr, 1974) should be made, i.e., because a detector has "fired" does not mean one can identify the cause of firing. It is obvious that the simple detector described above can be fooled, but is also true of more sophisticated detection algorithms. For example, Exhibit 4.5 shows all of the seismograms from event 116.1720. The vertical dotted lines denote the P arrival times. When the event is located using the arrival times shown, the arrivals at pnm and rdm are six seconds early. This suggests than a small increase in the noise level at these stations might have been picked rather than the actual first arrival. Unfortuantely, the only other obvious alternative is ten seconds later than these arrival, not six seconds. Actually, the correct interpretation is that the arrivals at pnm and rdm are P arrivals and all the others are actually S arrivals. When these arrivals are interpreted as S and the event is relocated, all of the arrivals are consistent. The interesting thing is that these arrivals were hand picked!

There are two fundamental reasons why an algorithm will make an error. The first is that the descriptive power of the algorithm is insufficient for it to distinguish between certain cases. This is obvious for the simple power-law detector described above, but is a universal problem which must be contended with by any algorithm. Secondly, the algorithm may not have enough information available for it to properly make a decision. The fact that even trained seismologists can mistake S arrivals for P arrivals clearly indicates the ambiguity involved in identifying seismic signals using information from only one sensor.

Information from many different levels must be used to correctly identify events. This is how a seismologist goes about the event analysis problem. The people ("readers") at NCER and Caltech who read seismograms on a daily basis have been observed during their work. The different types of information the readers utilized in picking arrivals is quite impressive. Arrivals were never picked without looking at most of the seismogram first. Very weak arrivals were picked after confirming that an event occurred on other traces. Subtle changes in signal characteristics such as a change in dominant frequency were used to pick low amplitude emergent arrivals. Familiarity with the operational characteristics of stations was also used. Some stations had notoriously bad residuals and arrivals picked from them were given lower weights in the event location. Once an event was located, arrivals which were inconsistent with this location were repicked.

The fact that an aftershock sequence was occurring was also used in picking. Since aftershocks for California earthquakes tend to have the same source characteristics, the signals from different shocks may be similar for the first cycle or so. Also, the relative arrival time between two stations may be roughly constant over several events. During an aftershock sequence, a number of multiple events might be expected to occur. It is important to distinguish between a seismogram consisting of the P and S envelopes from one event and a seismogram consisting of two overlapping envelopes from different events. This may be determined by checking if the supposed P and S envelopes have the same relative times on different traces. If they do, it is a multiple event.

The main thrust of this thesis is that to provide accurate arrival times and epicenters, an automatic system should use as much information from as many different sources as possible. One of the most powerful constraints is to use information on other sensors. Anderson (1976) showed that simple but robust heuristics using the spatial and temporal pattern of arrivals over the network could be used to improve the performance of the CDS system. His method correctly classified ninety four percent of the arrivals.

Since a tentative event location can provide the most narrow constraints on arrival times, the processes of picking arrivals and locating the event are combined in an iterative fashion. The arrival time and standard error of each arrival is determined by applying the algorithm presented in the next section to each seismogram. This information is then used to determine an initial location for the event using an algorithm which is insensitive to travel time errors and which is developed in Chapter 5.

Once a trial location is available, arrival time residuals can guide further processing. For example, a positive residual indicates that the arrival is late compared to what the location predicts. Thus if the station was between 120 and 160 km. from the event, this suggests that the algorithm may have picked the stronger mantle reflection, PMP, and should look for a weaker longer period arrival that might be a refracted arrival originally missed. A large negative residual means that the arrival is earlier than predicted, and might suggest that a glitch or a change in the noise level may precede the actual arrival. Of course, one should not rely on residuals too heavily since they can reflect more the inadequacies of the model than inaccuracy of the picks (Anderson, 1978). Arrivals inconsistent with the original location can be repicked appropriately and the location and picking process can be repeated as often as necessary.

Ideally, the algorithm has the potential of using partial information (a tentative location) to operate on the data in a way that could not be used otherwise. Using multiple levels of information makes processing more complicated. However, it provides constraints which are not available at the lower levels of the system. Using tentative picks to suggest a location allows partial knowledge to be used to control the analysis in a manner which is not irrevocable. That is, an hypothesis can be suggested and either accepted or rejected and the processing modified accordingly. Thus such an algorithm is truly data adaptive.

4.1. PICKING ALGORITHM

This section describes SAM, the picking algorithm used (See Exhibits 4.6 to 4.9). [1] The algorithm is somewhat similar to one developed by Stewart, (1977), which is being used by the CDS real time location system. Although real time operation is not required, the algorithm is sufficiently simple to allow its use in a real time envir^nment and it can even be implemented on a microprocessor (Anderson, 1978). The idea behind a simple algorithm was that the analysis of bugs in the algorithm would show how the algorithm could be improved. In fact, the bugs that were found are general enough that they must be contended with by any picking algorithm. Also, it was felt that work on a better algorithm should be postponed until the aliasing problem discussed in Chapter 2 had been resolved.

A minimum amount of prefiltering is done. Only a two second running dc average is removed from the data. Although additional filtering might enhance the signal somewhat, it was felt that since the signals can be clipped, any additional filtering might drastically distort the signal. Also, filtering alone is not likely to completely eliminate any one of the problems that the algorithm must face. Prefiltering can, of course, be added without modifying the algorithm.

The algorithm uses the interval of data between two consecutive zero crossings as its basic unit of data which will be refered to as a "blip." A blip is described by:

---

[1] Algorithms are presented in a pseudo computer language similar to RATFOR (Kernighan and Plauger, 1976) which is a structured algebraic language similar to ALGOL or PL1. Compound statements are contained in braces. Anything following a "#" is a comment.

t1blip: the time of the first zero crossing

t2blip: the time of the second zero crossing

tpeak: the time of the maximum amplitude between the zero crossings

peak: the maximum amplitude between the zero crossings

width: time between zero crossings (pulse width)

ttrig: the time at which the signal first exceeded the trigger threshold crossing (if it did)

xtrig: the amplitude of the first trigger threshold crossing

That is, the data between two zero crossings is represented by either three or four points. This description provides amplitude and frequency information used to distinguish between noise and seismic signals. The subroutine getblip(blip) is the interface between the the raw seismic data and the SAM routines. It scans the data returning either the next blip or an end of data indicator, eod. Getblip also determines a two second running average of the absolute value of the input, avabs. Avabs is used to estimate the background noise level and care is taken so that it is not contaminated by glitches or seismic signals.

Initially, CEDAR provides SAM with an estimate of the dc level and noise level for the channel. It then begins by scanning the input blip by blip until a trigger occurs. A trigger is defined by a blip whose peak exceeds the amplitude threshold, trigthr, and whose width is within certain limits. This blip is then considered a potential first arrival. (Reasonable values of trigthr range between three and six times avabs.)

Once a trigger occurs, SAMP and SAMC (detailed below) make semantic checks on the data following the trigger in an attempt to distinguish actual first arrivals from false triggers. These false triggers occur at a rate of about 100 times a second on the CDS (Stewart, personal communication 1976) and CEDAR systems. If SAMP is satisfied that a P arrival has occurred, SAMC collects the coda of the seismogram and makes further checks.

The algorithms of Stewart (1977) and Allen (1976) operate similarly. In their procedures, a decision about an arrival is put off until the data after it has been inspected. If this interval of data passed certain conditions the original onset is considered as genuine. If not, the interval is considered as noise and processing continued. This method can give rise to the following bug:

> Suppose a false trigger occurred just before an actual arrival, such that the actual arrival was contained in the inspection interval. If the inspection of the trigger rejected it as a first arrival, as it should, the actual arrival would be thrown away as well, since the entire inspection interval is never considered again.

A better approach is to back up to the point just after the false trigger and reconsider the data; this is what SAMP does. When a trigger occurs, the data following it is inspected. If based on this inspection, the trigger is not considered an arrival, the algorithm backs up to the blip following the one which caused the trigger and continues to look for an onset. In this way, the algorithm has the luxury of being able to pursue a hypothesis for a short time without really being committed to it. If the hypothesis is wrong, it can back up and begin afresh.

If the trigger meets the onset condition, SAMC attempts to collect the coda of the seismogram. If a coda is found, SAM returns a description of the seismogram. If not, the entire interval from the trigger to the end of the coda processing is considered to be a noise burst, and the search for an onset continues without backing up.

SAMP looks at the next tp seconds of data and backs up if the trigger was a glitch or if the input stays below the trigger threshold for too long a time. If the inspection interval is successfully examined, the signal is checked to be sure that its amplitude and dominant frequency [1] is reasonable.

SAMC's logic is similar to SAMP's except that SAMC does not backup. The end of the coda is reached when the input amplitude stays below the coda threshold, codathr, for more than two seconds. The requirement that the duration of the seismogram is at least four seconds is equivalent to requiring that the event is larger than magnitude 0.4 (Stewart, 1977; Lee, Bennett, and Meagher, 1972).

The algorithm is obviously very simple and will accept a large variety of signals. Although it computes the dominant frequency of the seismic signal, it does not compare it to the dominant frequency of the noise. Thus certain weak arrivals which can only be distinguished by a change in frequency could be missed. Also, it makes no use of waveshape

---

[1] The dominant frequency is taken to be 1/2 of the number of blips divided by the total duration of the blips.

or envelope information which could give valuable distance information. Unfortunately, although these heuristics are quite helpful for seismologists, they are rather difficult to implement. The variability of the arrivals is so great that if one tries to specify or predict the characteristics of the arrival or the envelope of the seismogram too much, the performance of the algorithm can be degraded rather than improved. After a trial location is available from the results of this simple picker, these heuristics should be easier to apply and can be used to improve the picks.

4.2. ESTIMATING ARRIVAL TIME ERROR

A number of theoretical expressions for the error of an arrival time of first motion of a seismic signal have been derived depending on the model of signal and noise used (Woodward, 1957 p. 105; Helstrom, 1960 p. 214; Mityashev, 1965 p. 14-16; Pakiser and Steinhart, 1964; Aki, 1976; and Torrieri, 1972). Following Helstrom (1960) and Woodward (1964), Capon and Green (1970) derived an equation for the standard deviation hopefully achievable in the estimation of arrival times:

$$\sigma_t = \frac{1}{B\sqrt{R}} \qquad \text{EQ 1}$$

Where $R = 2E/N$ where E is the total power of a known signal and N is the noise power per unit bandwidth. B is a measure of the spectral extent of the signal given by:

$$B^2 = \overline{w^2} - \overline{w}^2$$

Where

$$\overline{w^2} = \frac{1}{A}\int w^2 S(w)\,dw; \quad \overline{w} = \frac{1}{A}\int wS(w)\,dw; \quad A = \int S(w)\,dw$$

Where S(w) is the power spectrum of the signal. This equation shows that the error of the arrival is inversely proportional to both the signal bandwidth, B, and the square root of the signal to noise ratio. Since the amplitude and bandwidth of a seismic signal decrease with distance, the accuracy of the arrival time should decrease quite rapidly with distance as has already been shown in Chapter 3. Unfortunately, the shape of the seismic signal is unknown and the equation is not directly applicable. It is also difficult to decide what is a good measure of R since as Aki (1976) has pointed out, the amplitude of the first motion is zero by definition. Estimating R from the first few cycles of the signal may not be very good as the amplitude of the first half cycle can be quite a bit smaller than following cycles.

The algorithm for estimating the arrival accuracy is as follows: The first half cycle of the first arrival is represented by four points: the first zero crossing, the first trigger threshold crossing, the first maximum in the half cycle, and the next zero crossing as described above. The standard deviation of the arrival time is computed by:

$$\sigma_t = \frac{\sigma_n}{M} \qquad \text{EQ 2}$$

where σ_n is the standard deviation of the noise and M is the slope of the line between the threshold crossing and the maximum. Heuristically speaking, the equation says that for identical arrivals, the accuracy is determined by the noise level, and for the same noise level, sharper arrival should be more accurate than weaker ones. This agrees well with the normal practice of weighting impulsive arrivals higher than emergent ones in an earthquake location.

One of the difficulties of applying EQ 2 is how to estimate M. Four points do not always represent the shape of the first arrival adequately. The accuracy of very sharp arrivals is likely to be overestimated and weaker ones may be underestimated. This is because M is estimated near the top of the leading edge of the first cycle. For impulsive arrivals, this portion of the cycle is usually steeper than the actual moment of onset. Weaker arrivals tend to be more rounded than impulsive arrivals and M will usually be less than the slope of the onset. At least to begin with, this was considered desirable since it tended to clearly separate the accurate arrivals from the others.

4.3. EXHIBITS

4.1. A selection of seismograms from event 116.414 which is typical of the medium sized events recorded by CEDAR. Thirty seconds of data is shown, three seconds before the machine picked arrival time and twenty seven seconds after. Each seismogram is annotated as in Exhibit 1.1. wwv is shown in the last trace as a time scale. The leading edges of consecutive square waves are one second apart.

4.2. Examples of the three types of glitches occurring in the CEDAR data (see text). Thirty seconds of data are shown.

4.3. Examples of DC offsets and noise bursts which occur in the CEDAR data. Thirty seconds of data are shown.

4.4. Examples of crosstalk. a) sbb talking to mwc. The upper two traces are a blow up of the first 10 seconds of the lower two traces. b) blt talking to rdm. The upper two traces are a blow up of the first 10 seconds of the lower two traces. Each seismogram is annotated as in Exhibit 1.1.

4.5. Seismograms for event 116.1720. The vertical dotted line marks the hand picked P wave arrival time. In four out of the 6 cases, the S arrival was mistakenly identified as the P arrival. wwv is shown as a relative time scale.

4.6. The terms the picking algorithm uses to describe seismograms.

4.7. SAM scans the input data (using getblip) for a trigger threshold crossing. It then calls SAMP and SAMC to determine if a seismic signal is present.

4.8. SAMP examines a short interval of data to decide if it is a P arrival.

4.9. If SAMP finds a P onset, SAMC operates on the data until it finds the end of the coda. It then checks to see that the coda's frequency and duration is reasonable for a seismic signal.

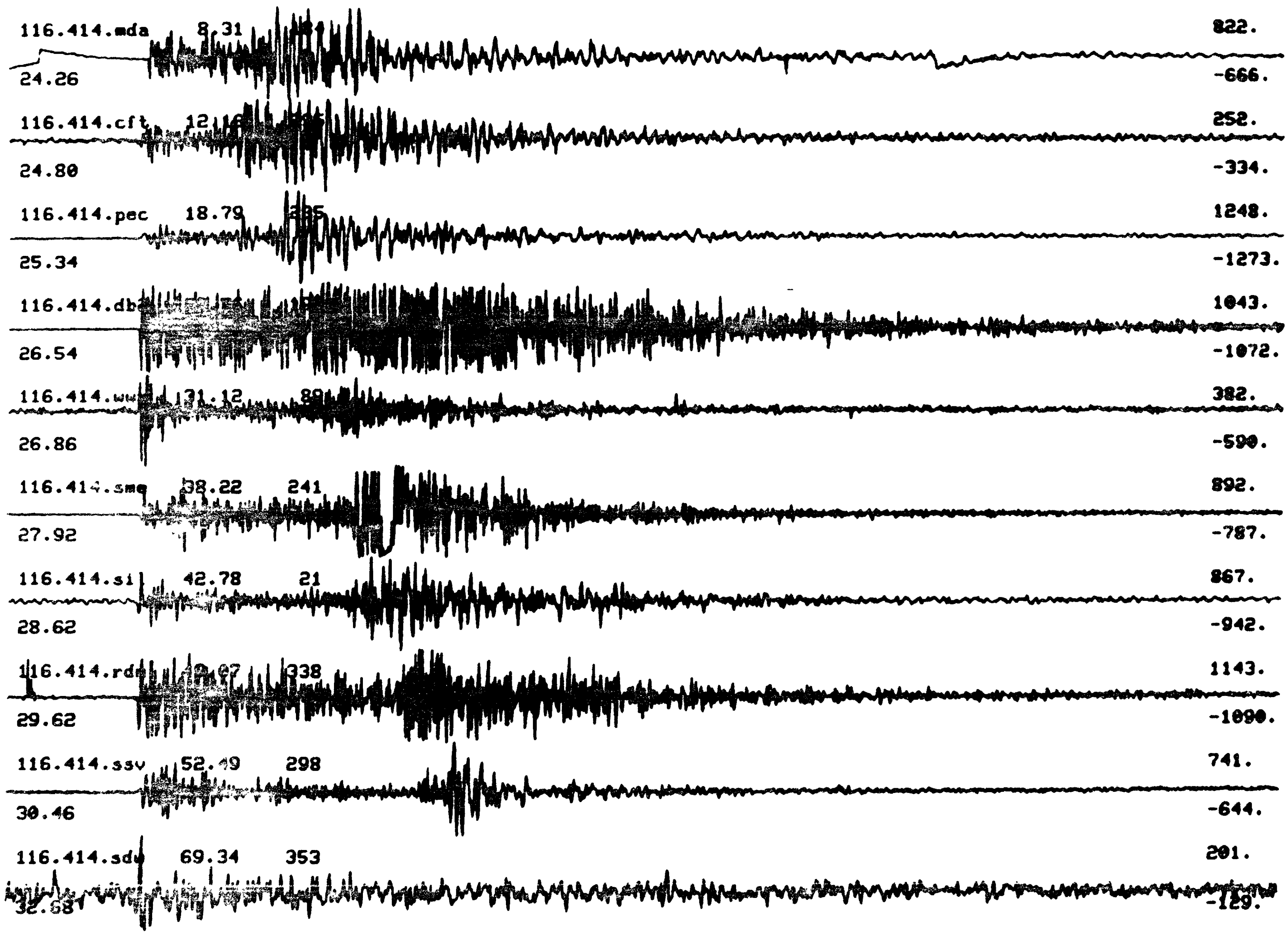

116.414.mda 8.31
24.26
822.
-666.
116.414.cft
24.80
252.
-334.
116.414.pec 18.79
25.34
1248.
-1273.
116.414.db
26.54
1043.
-1072.
116.414.ww 31.12
26.86
382.
-590.
116.414.sme 38.22 241
27.92
892.
-787.
116.414.si 42.78 21
28.62
867.
-942.
116.414.rd 338
29.62
1143.
-1090.
116.414.ssv 52.49 298
30.46
741.
-644.
116.414.sdw 69.34 353
201.
-129.

EXHIBIT 4.1

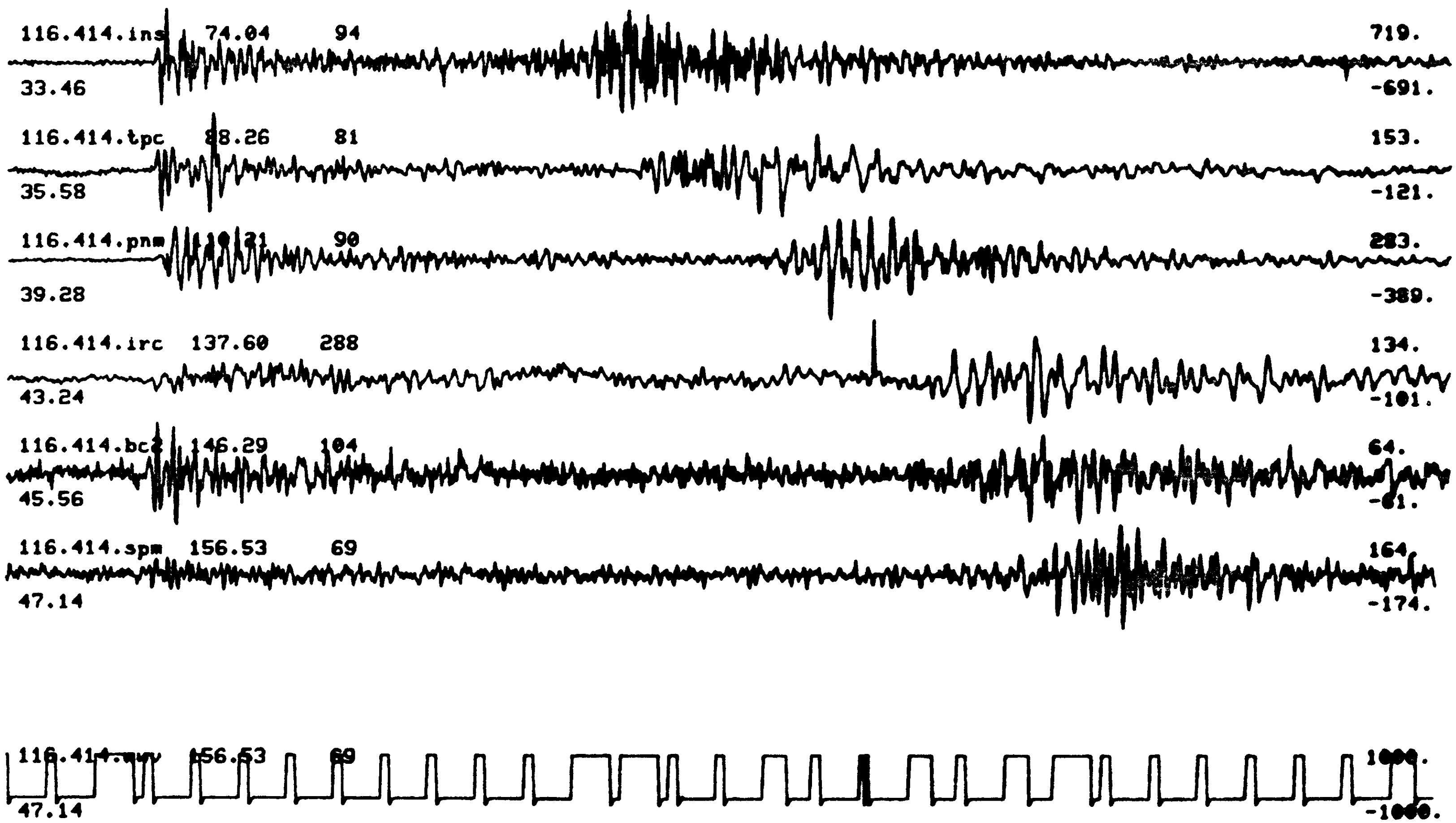


EXHIBIT 4.1 CONTINUED

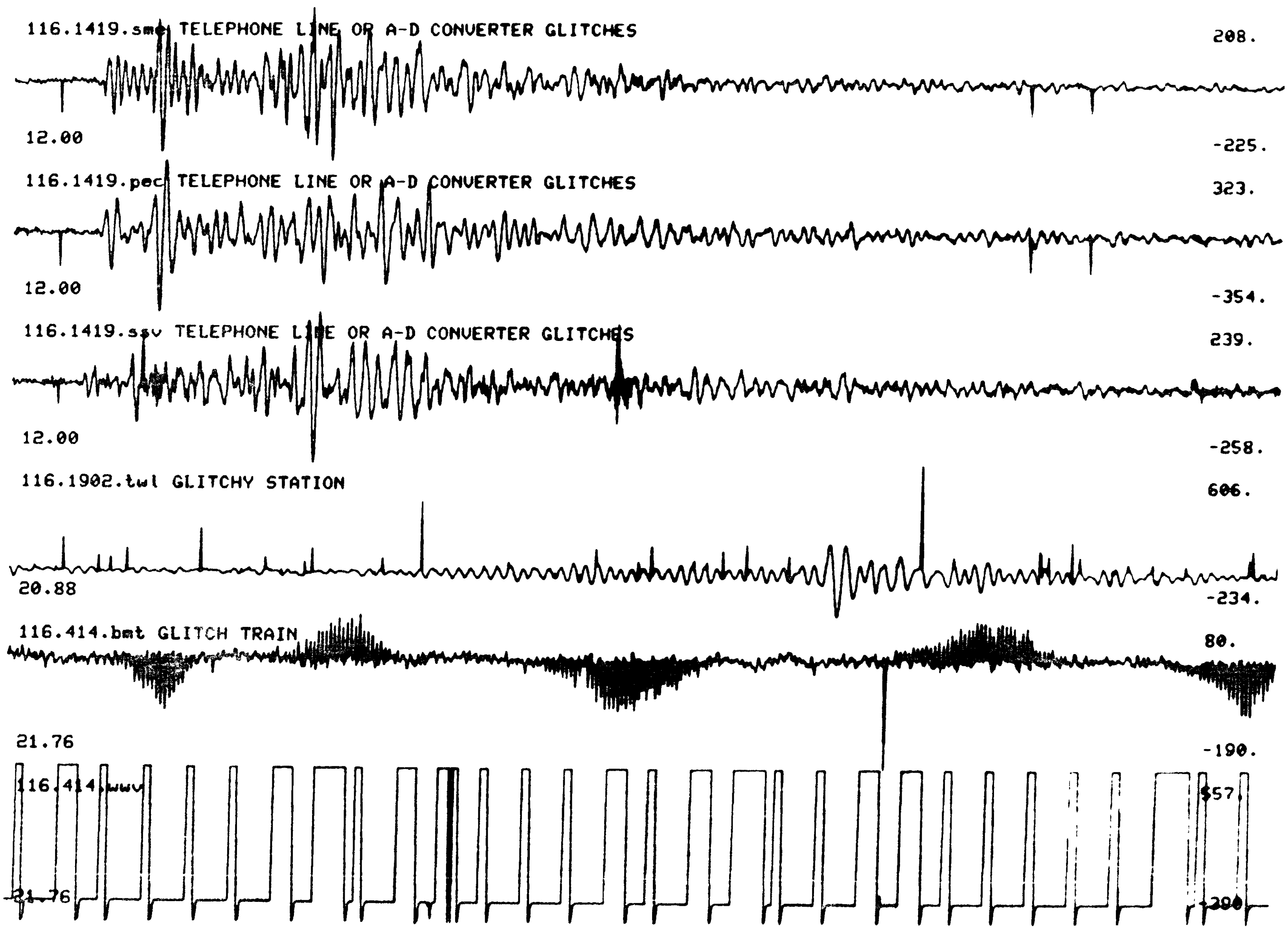

116.1419.sme TELEPHONE LINE OR A-D CONUERTER GLITCHES
208.
12.00
-225.
116.1419.pec TELEPHONE LINE OR A-D CONUERTER GLITCHES
323.
12.00
-354.
116.1419.ssv TELEPHONE LINE OR A-D CONUERTER GLITCHES
239.
12.00
-258.
116.1902.twl GLITCHY STATION
606.
20.88
-234.
116.414.bmt GLITCH TRAIN
80.
21.76
-190.
116.414.wwv
21.76
-200.

EXHIBIT 4.2

EXHIBIT 4.3

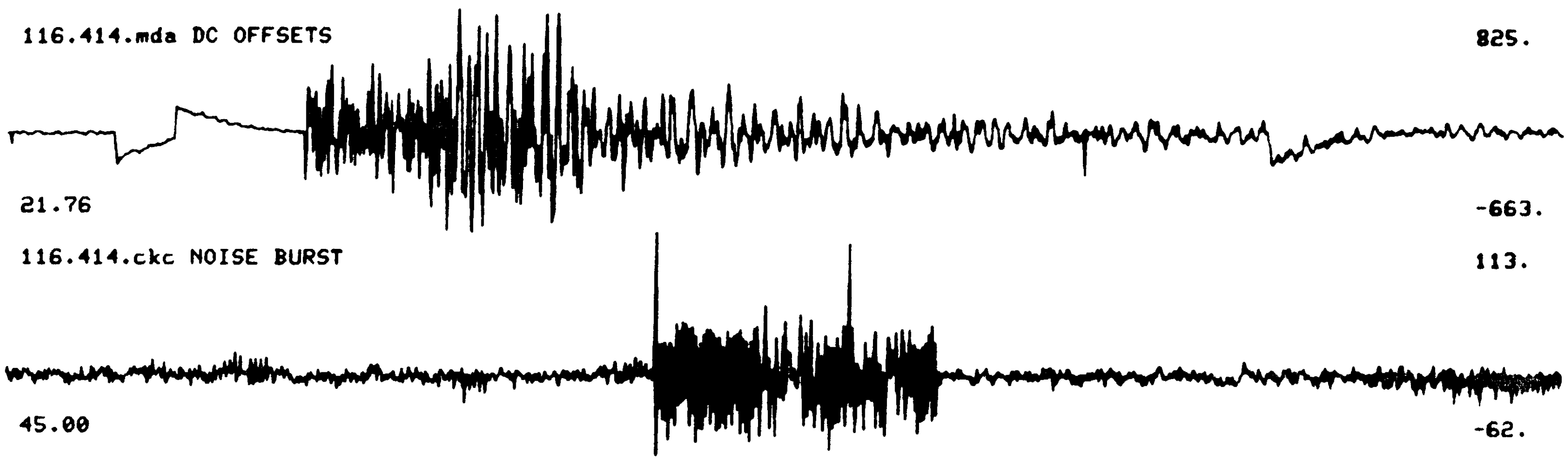
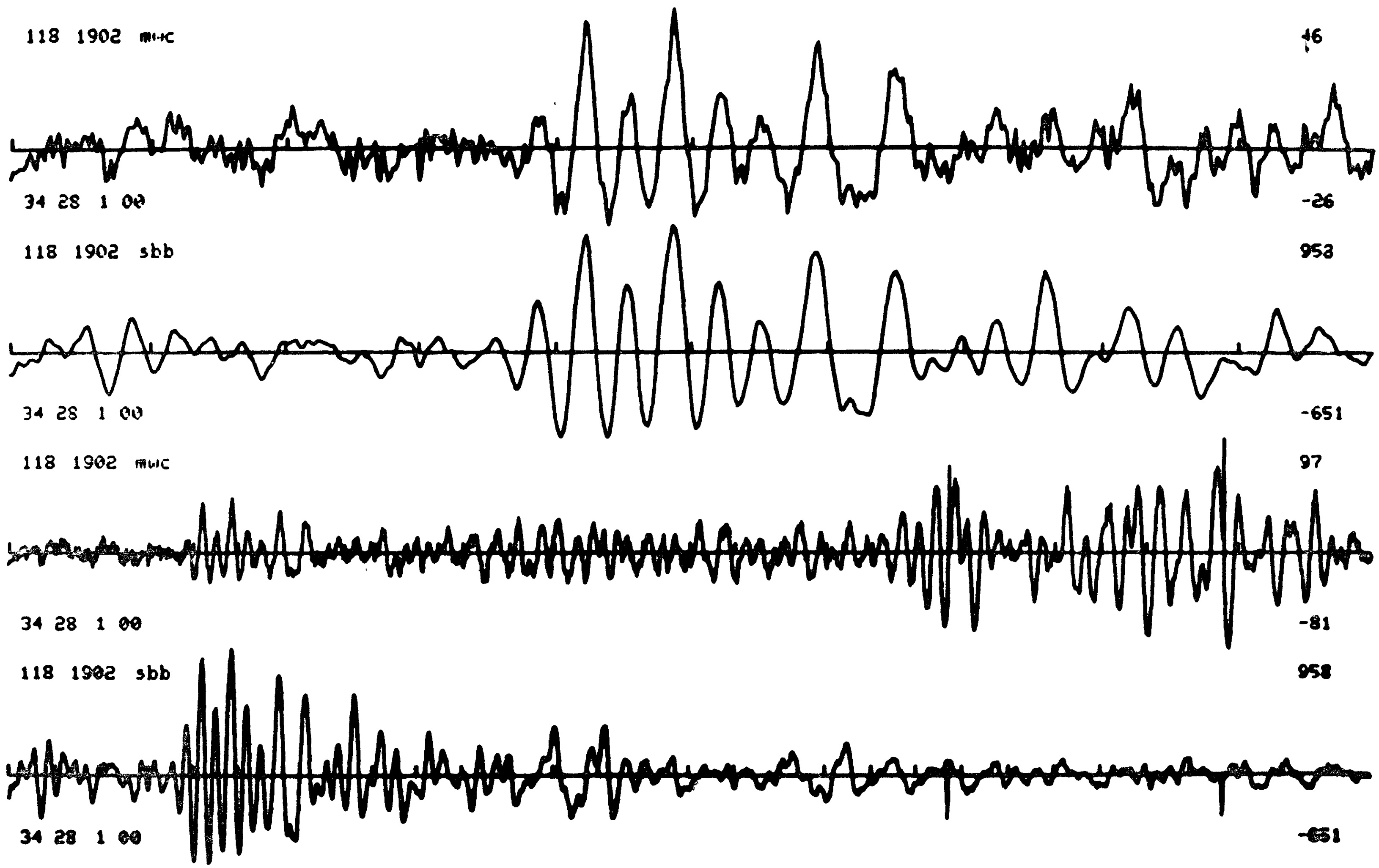
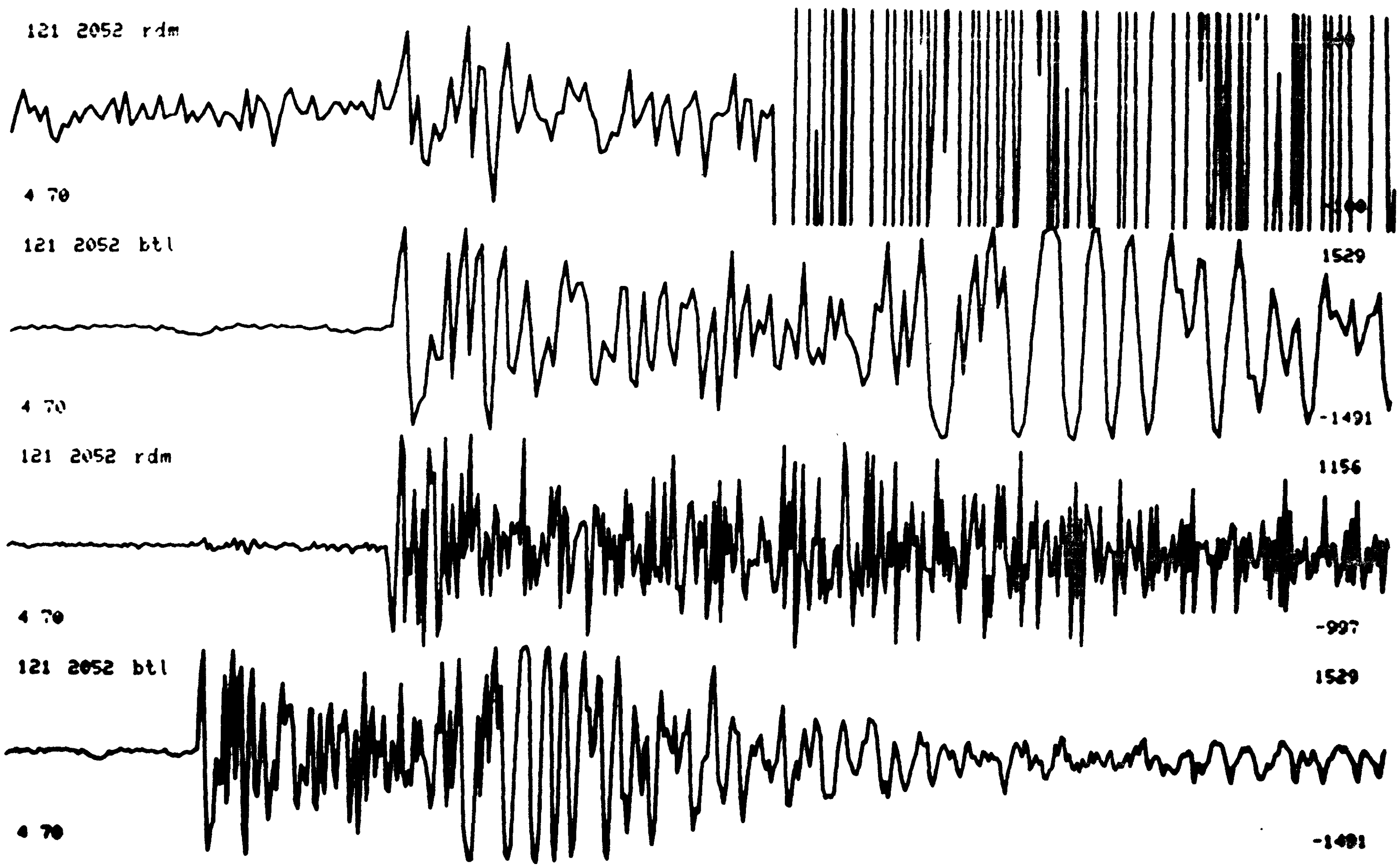

116.414.mda DC OFFSETS
825.
21.76
-663.
116.414.ckc NOISE BURST
113.
45.00
-62.

EXHIBIT 4.4-a



EXAMPLE OF CROSS TALK



EXAMPLE OF CROSS TALK

EXHIBIT 4.4-b

EXHIBIT 4.5

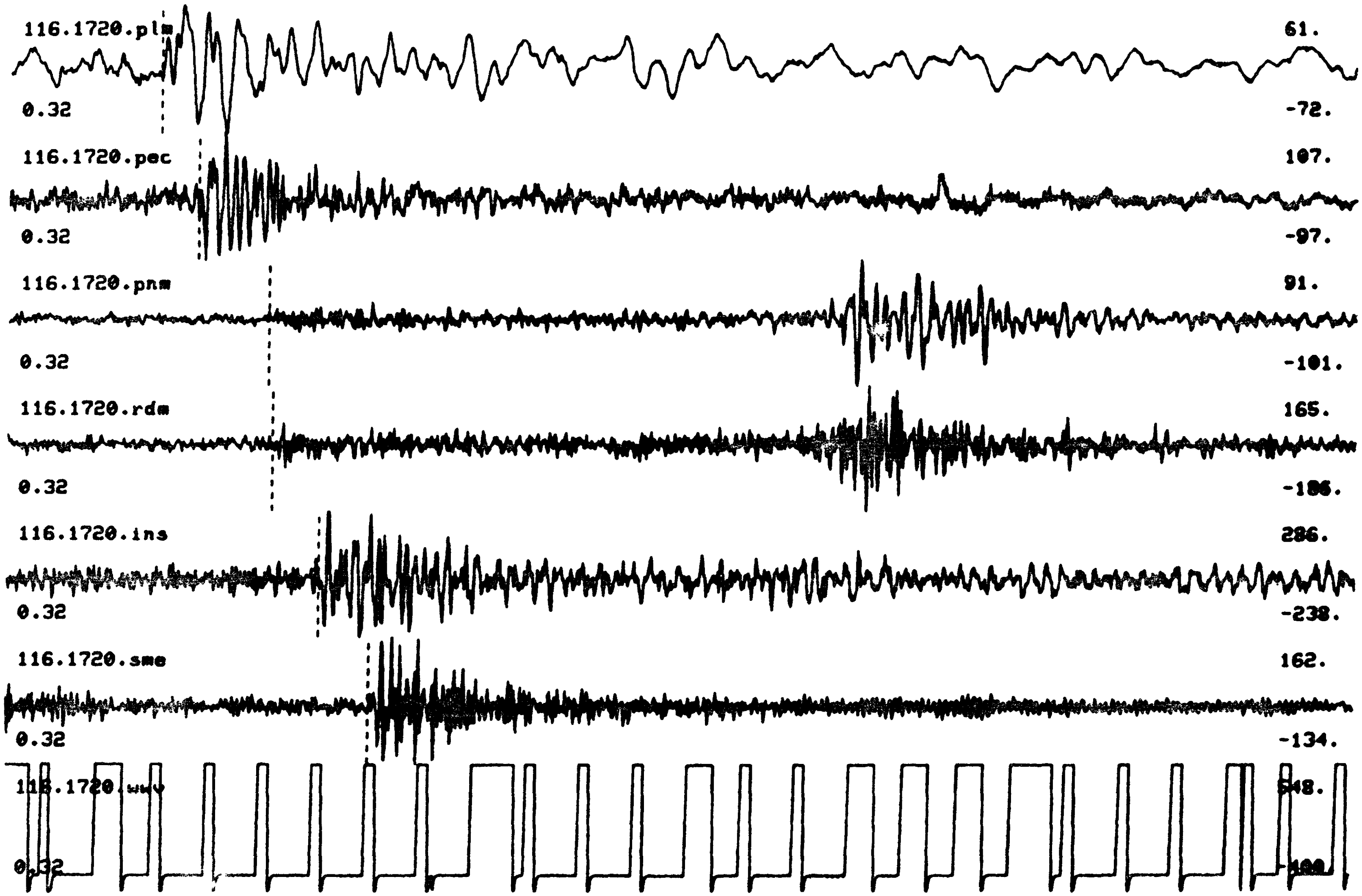

116.1720.plm
61.
0.32
-72.
116.1720.pec
107.
0.32
-97.
116.1720.pnm
91.
0.32
-101.
116.1720.rdm
165.
0.32
-186.
116.1720.ins
286.
0.32
-238.
116.1720.sme
162.
0.32
-134.

EXHIBIT 4.6

SEISMOGRAM DESCRIPTION DETERMINED BY SAM ROUTINES

noise description

dc	dc bias
avabs	average absolute value
ltrig	number of false triggers

p arrival description

ton	onset time
terr	error
pwidth	pulse width of first half cycle
motion	first motion amplitude
tpmax	time of maximum amplitude
pmax	maximum amplitude
avfrqp	average frequency
lhip	number of peaks > trigthr

coda description

fmp	time between ton and the end of coda (F-P time)
tcmax	time of maximum amplitude
cmax	maximum amplitude
avpeak	average peak height in coda
avfrqc	average frequency
lhic	number of peaks > codathr

OTHER TERMS USED BY SAM

dt	sampling period
trigthr	trigger threshold
codathr	coda threshold
eod	end of data indicator

EXHIBIT 4.7

```
function sam()
# sam:  returns yes if an arrival is found

  while(getblip(blip) != eod)
     if (narrow < width & width < wide &
         abs(peak) > trigthr &
         samp(blip) == yes &
         samc() == yes) return yes
  return no
  end
```

SYMBOL	TYPICAL VALUE	MEANING
narrow	0.0	width of onset must be > narrow (sec.)
wide	0.5	width of onset must be < wide (sec.)

EXHIBIT 4.8

```
function samp (blip)
# samp: P acceptor - returns yes if a P arrival found
#       starting at blip

  tfail = t2blip  # time to fail back to

  lhi = 0   # number of peaks > trigthr
  lblips = 0   # number of blips
  maxaft = 0  # largest amplitude after triggering blip

  while (getblip(blip) != eod & t2blip - tfail < tp) {
     lblips = lblips + 1

     # if amplitude just after onset < 1/4 first peak => glitch
     if (lblips > 1 & abs(peak) > maxaft) maxaft = abs(peak)
     if (lblips == 4 & maxaft < abs(motion)/4)
        fail

     if (abs(peak) > trigthr) {
        lhi = lhi + 1
        tlo = 0
        }
     else {
        tlo = tlo + width
        if (tlo > tlomax)   # amplitude low too long
           fail
        }
     }

  if (ampmax >minampp &
      lhi >= 3.0/tp &
      minavfrqp <= avfrqp & avfrqp <= maxavfrqp) {
     "compute P arrival description"
     return yes
     }
  else fail
  end
```

SYMBOL	TYPICAL VALUE	MEANING
tlomax	.25 - 3.0	maximum time amplitude can remain below trigthr (sec.)
minampp	10.0	minimum acceptable p amp. (digital units)
minavfrqp	0.5	minimum allowable p freq. (Hz.)
maxavfrqp	20.0	maximum allowable p freq. (Hz.)
tp	0.5 - 3.0	duration of inspection interval (sec.)
fail		backs up processing to time tfail

EXHIBIT 4.9

```
function samc ()
# samc: coda acceptor - returns yes if coda found

  lhi = 0    # number of peaks > codathr
  lblip = 1   # number of blips in coda

  while (getblip(blip) != eod) {
     lblip = lblip + 1
     if (abs(peak) > codathr) {
        lhi = lhi + 1
        tlo = 0
        }
     else {
        tlo = tlo + width
        if (tlo > 2.0) break    # end of event
        }
     }

  duration = t2blip - ton  # length of seismogram
  if ((width == eod ¦ duration > minduration & lhi/duration > 3 ) &
       avfrqc > minavfrqc & avfrqc < maxavfrqc &
       avpeak > 2.0*avabs) {
        "complete coda description"
        samc = yes
        }
  else   # reject coda as noise
     samc = no
  return
  end
```

SYMBOL	TYPICAL VALUE	MEANING
minavfrqc	0.5	minimum allowable coda frequency (Hz.)
maxavfrqc	20.0	maximum allowable coda frequency (Hz.)
minduration	4.0	minimum acceptable seismogram duration (sec.)

5. EXPLOITING NETWORK CONSTRAINTS

The previous chapters showed that arrival time can not always be determined unambiguously using only waveform information from a single sensor. Clearly, an estimate of the location of the event would provide the most useful information on which arrivals were misread. However, a reliable location can be obtained only if the procedure for estimating the location is not strongly affected by large errors in the travel time observations. In statistics, estimates which are resistant to larger errors are called "robust." Jeffreys first realized the need for robust methods in seismology in the 1930's (Jeffreys, 1962 p. 94-95) and his technique is still used today in locating earthquakes (Lee and Lahr, 1975; Bolt, 1960; ISC, 1976 p. iii). More recently, a great deal of research has been done on robust estimators in the statistical community. Most of the work has dealt with one-dimensional robust estimation (see Section 5.2.1) which is now fairly well understood (Andrews et al., 1972). Less work has been done on the more complicated problems of robust linear regression (Andrews, 1974; Beaton and Tukey, 1974) and robust non-linear regression (Dennis and Welch, 1976).

This chapter is concerned with the interrelated tasks of identifying bad arrivals and determining a robust hypocenter. The first section presents a new method for eliminating large errors which does not require a location estimate. The following sections introduces some of the important ideas in robustness theory. The third section addresses the problem of robust earthquake location. A new method for determining an initial location is presented. Several location methods are compared, and the effect of arrival time error on the location determined by least squares and a robust method is investigated.

5.1. ARRIVAL PAIR CONSISTENCY

The time of an arrival at one station determines an upper and lower bound on the time of arrival at any other station in the network. This is because the difference in arrival time between two stations should be less than or equal to the time it would take for a seismic wave to travel from one station to another. Expressed as an inequality, this is given by

$$|T_i - T_j| \leq t(i,j) \qquad \text{EQ 1}$$

Where T_i and T_j are the observed arrival times at stations i and j; $t(i,j)$ is the travel time between stations i and j. The inequality is most useful for a dense network in which the distance between two stations is expected to be small and there is a larger chance of a pair of stations being roughly colinear with the epicenter.

When any pair of arrivals fails to satisfy EQ 1 , one or both of the arrivals is likely to be in error. The problem is to try to identify which arrival is in error. Hopefully, a bad arrival will be inconsistent with more than one other station and it can be identified using the following algorithm:

Make a list of all the station pairs which fail to satisfy EQ 1. The station which occurs most often in this list is the most likely to be in error. All pairs in the list containing this station can be removed from the list. The station occurring most often in the resulting list is then the second most inconsistent arrival and its pairs can be removed similarly. This process can be repeated until the list is empty or no station occurs in the list more than once. We cannot determine which arrival of the remaining pairs is in error using only EQ 1. [1]

This algorithm is used as a coarse sieve which removes large errors before a location is determined. The median nearest-neighbor distance for the CEDAR array is 20 km. The median second-nearest-neighbor distance is 24 km., and 75 percent of the stations have a second-nearest-neighbor distance of 34 km or less. Thus considering the fact that the station density is higher than this in regions of the net with most seismic activity, the sieve will usually detect errors of 3.0 sec. or more and can detect smaller errors in certain cases.

The algorithm identifies 20 bad arrivals in the handpicked data set consisting of 1190 P wave arrivals from 61 events. The bad arrivals detected had residuals which ranged from -15 to +11 sec.; the smallest absolute residual identified was 1.6 sec. The large positive residuals (late arrivals) are usually S arrivals on distant (> 100 km.) Benioff stations. They are usually given low quality weights by an analyst. At least some of the large negative residuals are due to cross talk (see Chapter 4). and are given quality weights which may be quite high.

When used with machine picked arrivals, the consistency sieve may identify several larger errors for one event (see Chapter 6). As with the hand picked arrivals, large positive errors are usually S arrivals on Benioff stations. Most human operators would not bother with such weak arrivals, but the algorithm is currently set to return very weak

---

[1] It cannot be guaranteed that a station which is inconsistent with several other stations is actually the bad one. It may be that it is a good station and the others are bad. However, it can be argued on probabilistic grounds that this is less likely to occur.

signals and is thus prone to these types of errors. Early miss-picks are either due to cross-talk or due to small changes in noise level in front of the actual first arrival which is interpreted as a weak first arrival such as Pn.

5.2. INTRODUCTION TO ROBUST ESTIMATION

The following sections will be concerned with determining a robust estimate of an earthquake hypocenter from a set of observed travel times. Unfortunately, this will require delving into the details of solving robust nonlinear regression problems which are much less understood than the related problems of either robust linear regression or nonlinear least squares. A summary of robust estimation will be presented in the following three sections. Section 5.2.1 introduces the basic ideas of robust estimation. Section 5.2.2 deals with solving robust linear regression problems since their solution is required to solve the nonlinear robust regression problem discussed in Section 5.2.3. (See (Andrews et al.,1972; Huber, 1972; Andrews, 1974; and Hill, 1977) for a more detailed treatment of robustness.)

5.2.1. ONE-DIMENSIONAL ROBUST ESTIMATION

Given a set of observations, y_i, i = ,...,n, one is often interested in summarizing this set by two parameters, θ and s which describe the central tendency (location or mean) and spread (scale or variance) of the observations. Often, θ and s are considered as estimates of the parameters μ and σ, of an underlying probability density function.

Huber (1972) considered a class of estimates of θ which may be found by minimizing:

$$P = \min_{\theta} \sum_{i=1}^{n} \rho\left(\frac{y_i - \theta}{s}\right) \qquad \text{EQ 2}$$

or equivalently by finding the θ which satisfies

$$\sum_{i=1}^{n} q\left(\frac{y_i - \theta}{s}\right) = 0. \qquad \text{EQ 3}$$

where $p(e)$ is a penalty function, $q(e) = p'(e)$, and s is a scale estimate.

If the observations are assumed to come from an underlying probability density function, $f(\frac{y - \mu}{\sigma})$, then the solution of EQ 3 is a maximum likelihood estimate of μ if $q = f'/f$ and $s = \sigma$ (Jeffreys, 1948). [1] For example, if the data is assumed to follow a Gaussian distribution, the $p_G(e) = e^2$, $q_G(e) = e$, and $\theta = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Location estimates which are not affected by a few larger errors in the data and which are good for a large family of underlying distributions are considered robust. The sample mean is not robust since a single larger error can move θ arbitrarily far from the true μ.

The form of $q(e)$ which Huber considered is [2]

$$q_H(e) = \begin{cases} e & |e| \leq k \\ -k & t < -k \\ k & t > k \end{cases} \qquad k = 1.345 \qquad \text{EQ 4}$$

This corresponds to a quadratic loss function for errors less than k and a linear one for larger errors. Two other estimators in use are the

---

[1] Because of this similarity to maximum likelihood estimates, Huber (1972) refers to such estimates as M-estimates. Several other forms of robust estimators have also been studied (Andrews et al., 1972; Huber, 1972)".

[2] The value of k given for the different estimators result in 95 percent asymptotic efficiency for the Gaussian distribution (Hill, 1977).

bisquare estimator (Beaton and Tukey, 1974)

$$q_B(e)=\begin{cases} e(1 - \frac{e^2}{k^2})^2 & \text{if } |e| < k \\ 0 & \text{otherwise} \end{cases} \qquad k = 4.685$$

EQ 5

and the sine estimator (Andrews, _et al_. 1972; Andrews, 1974)

$$q_S(e) = \begin{cases} k \sin(\frac{e}{k}) & \text{if } |e| < \pi k \\ 0 & \text{otherwise} \end{cases} \qquad k = 1.339$$

EQ 6

Jeffreys rocognized the need for robust estimation during his work on seismic travel times in the 1930's (Jeffreys, 1932, 1948, 1962, 1973). He showed that teleseismic travel time residuals seem to come from a Gaussian distribution contaminated by a small amount of a slowly varying background distribution, $g(r)$:

$$f(r) = ag(r) + \frac{(1-a)}{s(2\pi)^{1/2}}\exp(\frac{-r^2}{2s^2})$$

EQ 7

Where a is the amount of contamination, and $g(r)$ is fairly constant for $|r|$ < several s and $\int g(r)\,dr = 1$. This leads to a $q(e)$ of

$$q_j(e) = e\,[1 + c \exp(\frac{e^2}{2})]^{-1}$$

EQ 8

Where $c = \frac{s(2\pi)^{1/2}ag(r)}{1 - a} \sim$ constant. If there are enough observations, c may be estimated by dividing the frequency, $ag(r)$, of large residuals by the difference between ab and the frequency of residuals near the mode (Bolt, 1960); but in general, it must be estimated iteratively along with s and θ (Jeffreys, 1932). (In the following discussion, c will be assumed known.)

Given an initial θ, θ_0 and a corresponding scale estimate, EQ 3 can be solved iteratively, with the following iteration step: Let $w(r) = q(r/s)/r$, where $r = y - \theta_0$. Then EQ 3 can be written as

$$\sum_{i=1}^{n} w(r_i)(y_i - \theta) = 0$$

and a new estimate of θ is

$$\theta = \frac{\sum_{i=1}^{n} w(r_i)y_i}{\sum_{i=1}^{n} w(r_i)}$$

Thus in each step, the new θ is a weighted average of the observations. The weights are nearly constant for small residuals and decrease as the residuals increase. The effect of larger residuals will be less than in the least squares case (constant weights), and the estimate will be more robust.

Alternatively, the Newton-Raphson iteration step,

$$\theta = \theta_0 + \frac{s \sum q(\frac{r_i}{s})}{\sum q'(\frac{r_i}{s})}$$

may be used.

Andrews et al. (1972) studied 68 location estimators, and it is worth considering a few of their conclusions (Hill, 1977). Since the result of the above estimators is critically affected by the scale estimate, s, a robust estimate of scale such as

$$s = a \text{ median } \{|r_i - \theta|\}$$

should be used. Where a is chosen to make s a consistent estimate of the standard deviation if the observations came from the normal distribution (a = 1/.6745 (Hill, 1977)).

M-estimators perform very well for a variety of distributions. The best M-estimators seem to be those which redescend, i.e. $q(e) \to 0$ as $e \to \infty$ such as q_B, q_S, and q_J. Unfortunately, for the redescending estimators, EQ 2 does not necessarily have a unique minimum, and thus the final θ will depend on θ_0. Luckily, the initial estimate, $\theta_0 = \text{median}\{y_i\}$ has been found to be relatively good. In fact, M-estimates which take only one iteration step past the median are usually very good.

5.2.2. ROBUST LINEAR REGRESSION

This section is a brief summary of robust linear regression. We will be interested in finding the m by 1 solution vector ,x, to the matrix equation

$$y = Ax + r \qquad \text{EQ 9}$$

Where y is an n by 1 vector of observations, r is an n by 1 vector of residuals (errors), and A is an n by m matrix of independent variables. A robust estimate of x is one which minimizes with respect to x

$$\min_x \sum_{i=1}^{n} p\left(\frac{r_i}{s}\right) \qquad \text{EQ 10}$$

Where p(e) is one of the robust loss functions, $r_i = y_i - a_i x$ is the i'th element of r, $a_i = (a_{i1},\ldots,a_{im})$ is the i'th row of the matrix A, and s is a scale estimate for the residuals. Minimizing the above equation is the same as solving

$$\sum_{i=1}^{n} a_{ij} q(\frac{r_i}{s}) = 0 \qquad j = 1,\ldots,m$$

EQ 11

where $q(e) = p'(e)$. In matrix notation EQ 11 becomes

$$A^T q = 0$$

EQ 11

where q is an n by 1 matrix the i'th element of which is $q(\frac{r_i}{s})$. If $q = r$, EQ 11 reduces to the "normal equations" for the least squares solution of EQ 10, namely $A^T y = A^T Ax$

The robust regression problem EQ 11 can be transformed into a non-linear weighted least squares problem (Andrews, 1974; Hill, 1977). Let $q = Wr$, where W is an n by n diagonal matrix with diagonal elements $w_{ii} = q(\frac{r_i}{s})/r_i$. Then EQ 11 becomes

$$A^T W r = 0$$

Using $r = y - Ax$, this becomes

$$A^T W y = A^T W A x$$

EQ 12

If W were not a function of x, this would be a weighted least squares problem. EQ 12 can usually be solved by iteration. Both s and W may be computed from the previous solution, x, and a new solution is given by:

$$x = (A^T W A)^{-1} A^T W y$$

Although the convergence properties of this iterative scheme are such that it may be used with success (Hill, 1977) , the number of iterations required for convergence may be quite large depending on the type of $q(e)$ chosen (Armstrong and Frome, 1976) .

As with the one dimensional case, the scale estimate can affect the results of the regression. Hill (1977) recommends that

$$s = \frac{1}{.6745} \text{median}\{|y_i - a_i x|\}$$

be used as the scale estimate for the next iteration. Since the redescending estimators do not necessarily have unique solutions, the initial estimate of the solution, x_0, is crucial. One possibility is to use the least squares estimate of x as an initial solution. However, if the data errors are far from Gaussian, the least squares estimate may be biased away from the global maximum and a distant, local, maximum may be found (Andrews,1974) .

In the one dimensional case, the median was used as the starting value. This suggests that an estimate that minimizes the sum of the absolute value of the residuals (the L_1 estimate) be used as a starting estimate. Claerbout and Muir (1973) have recently publicized the robust properties of the L_1 norm for geophysical applications. They developed an efficient method of solving weighted L_1 problems of the form:

$$\min_x \sum_{i=1}^{n} w_i |y_i - a_i x| \qquad \text{EQ 13}$$

Where w_i are the weights: [1] Their method could be used to provide a starting solution for the weighted least squares iterations described above. The L_1 estimate should be a good initial estimate since it is much more robust than the least squares estimate and it has the advantage of not requiring an estimate of scale.

---

[1] Actually, they solve a more general problem, called the "skew norm problem", in which w_i depends on the sign of the residual.

5.2.3. ROBUST NONLINEAR REGRESSION

Given a set of observations $y_i, i = 1,\ldots,n$ we want to find a solution vector, θ, which minimizes

$$P(\theta) = \min_{\theta} \sum_{i=1}^{n} p(e_i)$$

EQ 14

where $p(e)$ is one of the robust penalty functions, $e_i = \frac{y_i - f_i(\theta)}{s(\theta)}$, where $f_i(\theta)$ is the theoretical value of the i'th observation, and s is a robust scale estimate.

Following the Gauss-Newton method of solving EQ 14, expand e_i in a first order Taylor series about a trial solution θ_0:

$$e_i = e_i(\theta_0) + \Delta e_i d\theta$$

EQ 15

Where $d\theta = \theta - \theta_0$, and

$$\Delta e_i = \frac{\partial e_i(\theta_0)}{\partial\theta} = \frac{-\partial f_i(\theta_0)}{s\ \partial\theta} - \frac{r_i \partial s(\theta_0)}{s^2 \partial\theta}$$

EQ 16

Since s should be fairly constant in the vicinity of the solution, the second term of the right hand side of EQ 16 is usually ignored. Plugging this equation into EQ 14:

$$P(\theta) \sim \min_{\theta} \sum_{i=1}^{n} p(e_i(\theta_0) + \Delta e_i d\theta)$$

EQ 17

Letting $y_i = e_i(\theta)$, $a_i = -\Delta e_i$, and $x = d\theta$, EQ 17 may be recognized as a robust linear regression which may be solved for the adjustment vector $d\theta$ as discussed in the previous section. The practical aspects of solving nonlinear robust regression applied to earthquake location will be discussed in the next section.

5.3. ROBUST EARTHQUAKE LOCATION

Before roughly 1960, graphical methods were often used for locating earthquakes (Ben-Menahem and Bath, 1960; Husebye, 1966). Since then, with the advent of computers, Geiger's method (which is essentially the method described in Section 5.2.3 using the least squares loss function) or a variation of it have been used quite commonly (Geiger, 1910; Bolt, 1960; Flinn, 1960; Nordquist, 1962; Engdahl and Gunst, 1966). A great deal of work has been done on problems involved in locating earthquakes using Geiger's method such as numerical stability (James et al., 1969; Buland, 1976; Smith, 1976; Bolt, 1970), convergence properties (Gershanik, 1973; Gutdeutsch and Araic, 1977; Buland, 1976), effects of station - epicenter geometry (Northrop, 1970; Peters and Crosson, 1972; Buland 1976), the effects of variation in the earth's structure (Greensfelder, 1965; Engdahl and Lee, 1976; Lee and Lahr, 1975; Gutdeutsch and Aric, 1977), and confidence limits (Flinn, 1965; Evernden, 1969). Unfortunately, all of this work is based on least squares, and since Jeffrey's work, little has been done on the problem of robust earthquake location (Bolt, 1976).

Recently, alternatives to Geiger's method have been proposed. Lomnitz (1977a) has proposed a method which determines the epicentral parameters separately from the focal depth and origin time. Although for the most part Lomnitz's method appears to be similar to Geiger's method (Smith, 1978; Lomnitz, 1978), an interesting aspect of his method is that the least squares minimization is constrained so that the mean slope in a plot of travel time residuals versus distance is zero. Garza, Lomnitz, and Ruíz de Velasco (1977) use an interactive computer program to determine epicenters graphically. Using triads of stations from a relatively small distance range their method can locate earthquakes reasonably well assuming that only the average seismic velocity is relatively constant over each triad.

Several proposals have been made to linearize the location problem in a constant velocity medium (Cisternas and Jobert, 1977; Miyamura, 1960· Miller and Harding, 1972). Although these methods produce locations quite rapidly they are not reliable because of numerical instability and robustness problems.

Exhibit 5.1 shows the location algorithm, SEEK, used in this thesis. It is a damped Gauss-Newton algorithm similar to that of Buland (1976). Although there are other attractive methods for solving this problem such as Marquardt's method (Marquardt, 1963, 1970, 1974), or more general optimization algorithms (Welsch and Becker, 1975; Dennis and Welsch, 1976), SEEK can be used with both the L_1 and robust weighted L_2 norms and it has been analyzed in detail for the L_2 case. Also Welsch and Becker (1975) have pointed out that it is not clear how iterating the weight matrix, W, and the scale, s, will affect Marquardt's algorithm.

The main difference between SEEK and Buland's algorithm is that in his algorithm the matrix W is fixed and the $P(\theta)$ is the least squares loss function. The while loop in Exhibit 5.1 reduces the size of $d\theta$ until a location which reduces $P(\theta)$ is found or $d\theta$ becomes so small that it is effectively zero. Although this does not guarantee a minimum, it works well in practice.

Clearly most of the problems are hidden in solving for $d\theta$. Buland (1976) has shown that in practice the matrix A^TA is often poorly conditioned, and he recommended using the QR algorithm when solving for $d\theta$. However, Marquardt (1963) and Smith (1976) show that proper scaling of the A matrix prolongs its numerical stability enough so that any reasonable algorithm for solving sets of linear equations can usually be used. However, since the robust case the W matrix is a complex function of θ through particular loss function and scale used, the numerical properties of A^TWA may at times be worse than for the least squares case. Even if A^TA is well conditioned, A^TWA may be very ill-conditioned and the undamped adjustment vector may be very poor. This can happen, as is shown in Section 5.3.2, when there are low weights on observations which contain all of the information about a parameter. Even if A^TWA is well behaved at a local minimum, a bad starting location could lead to bad results (Welsch and Becker, 1975); A further complication is that redescending robust loss function, which appear to be more robust than the non-redecending ones, need not have unique solutions.

Thus, although robust methods can generally provide better locations than least squares, one must be prepared for some pitfalls along the way. Clearly, the robust solution will be critically dependent on the starting location used. Traditionally, the first station to report an arrival is used as the trial location. Although Buland has shown that this is a good trial location for Gaussianly distributed arrival time errors, it is not particularly robust. In linear robust regression, the L_1 solution provides a reasonable starting solution. However, the L_1 norm is not always more robust than the L_2 norm for the event location problem (See section 5.3.3.). Because of the significance of the initial location, a new method of determining an initial location will be presented in the next section. It has been shown to be relatively robust and usually gives an initial epicenter which is within several km. of the final one. In Section 5.3.2, a comparison of several robust location methods will be made using the hand picked arrival times from CEDAR. A detailed comparison of the effect of errors on the robust and least squares locations will be made in Section 5.3.3.

5.3.1. ARRIVAL ORDER LOCATION METHOD

A new method which uses the order in which arrivals occur at different stations to determine a trial epicenter has been developed to provide a reasonably good starting location. Since only the arrival order is used and not the actual arrival times, grossly inconsistent arrivals will not affect the location. In the case of the CEDAR data, the median distance from the final epicenter is 2 km. so the method provides a good initial location for any robust location method. One interesting feature of the method is that it does not depend on a crustal velocity model.

For explanatory purposes, let us use the five station network shown in Exhibit 5.2. To simplify the discussion, the problem will be restricted to the latitude and longitude plane and for the moment, arrival times will have no error.

Using the first station to report the event as the initial epicenter uses one piece of order information. The idea behind the arrival order method is to use all the order information between station pairs to determine the smallest region containing the epicenter. A point on the edge of this region is taken as the trial epicenter.

The arrival order between each pair of stations provides a constraint equation. Given that the event is reported on station 2 before station 3 constrains the epicenter to occur somewhere to the north of the perpendicular bisector between the two stations (labled "2 / 3" in Exhibit 5.2). Similarly, each pair of stations constrains the epicenter to lie on one side of their perpendicular bisector or the other (see

Exhibit 5.2). If there are N stations, there are a total of N(N-1)/2 such constraints. [1] Since, each constraint is an inequality, this is essentially a linear programming problem. The region which satisfies as many of these constraints as possible is called the "feasible region" (Hadley, 1962). Rather than solving the problem as a linear programming problem, it was solved as an equivalent skew-norm problem (Claerbout and Muir, 1973). This is so that a skew-norm program could be used to provide both the arrival order location and an L_1 location for the event. The arrival order location determined by the skew-norm algorithm is an intersection of two of the perpendicular bisectors which satisfies as many of the order constraints as possible. For example, the arrow in Exhibit 5.2 points toward the arrival order location corresponding to the epicenter indicated by the star.

The set of bisectors partitions the plane into a set of convex regions. The order in which the stations report arrivals from an epicenter in that region can be used to label the region. For example, the star in Exhibit 5.2 occurs in region 52341. The labels of several other regions are shown in the Exhibit. Thus, given a set of arrivals one can immediately determine what region the event occurred in. The larger the number of stations, the smaller the average region size becomes. Exhibit 5.3 shows that for twenty stations the number of regions is quite large. The size of the regions is generally quite small in the central portion of the network, and increases toward the

[1] Actually most of these constraints are redundant. For data with no errors, only N constraints are necessary to determine a location. However, a redundant set of constraints is recommended when using real data. Using 2N or 3N constraints seems to give satisfactory locations in practice.

edge of the network. Outside the array, the regions become quite elongated. Thus the resolution of the arrival order location method is best in the center of the array and decreases toward the edges of the array. This appears to be a feature of earthquake location in general (Peters and Crosson, 1972).

Small errors will not affect the order of arrivals and thus will not affect the location. Larger errors will affect the order only if they happen to arrivals that are near each other in time. For example, if an error in the arrival from station 2 caused the arrival order to become 53241 then the event would appear to have occurred in the region just to the south of region 52341. (It turns out that the arrival order location does not change in this case however.) Gross errors correspond to region labels which do not exist. For example, if station S1 reported the earliest arrival by mistake, then the arrival order would be 15234. However, there is no region with this label since the arrivals are inconsistent. In such cases the most consistent feasible region is determined, which is 5234. Thus the redundant set of order constraints reduces the effect of inconsistent arrivals.

The only assumption made about the velocity structure is that it is laterally homogeneous. Although this is certainly not true for the real earth, it is about the weakest assumption that can be made since most location methods assume a specific laterally homogeneous velocity structure. The redundancy built into the method however, usually reduces the effect of actual lateral variations.

The arrival order location was used as the initial location for events using handpicked P wave arrivals. The arrival order results were compared with the final location to determine if they were better than the traditional method of using only the first station as the initial location. The results are shown in Exhibit 5.4. The median distance between the closest station and the epicenter was 11 km. The median distance from the arrival order location was 2 km and is almost always less than 3 km.

The influence of the number of stations is shown in Exhibit 5.5. If there are 10 or more arrivals, the arrival order location will usually be within 3 km of the final epicenter. This is more than the minimum number of arrivals (3) needed to locate the epicenter by Geiger's method. Essentially, since the arrival order method uses less information from each arrival, more arrivals are needed to get a well constrained location. However, the robustness of using less than say seven arrivals is extremely doubtful (see Section 5.3.3.). The interesting thing is that given a dense enough network, the arrival order method can provide reasonable epicenters without knowing the velocity structure under the network.

Thus, the arrival order location is recommended as the initial epicenter in any robust location method. The cost of computing the arrival order location is about the same as one iteration of Geiger's method, and usually reduces the number of iterations required to locate the event.

5.3.2. COMPARISON OF LOCATION METHODS

A program to test the various robust alternatives for earthquake location was developed. The program was designed so that a large number of initial locations and weighting schemes could be tried easily by specifying an ordered list of operations to be performed to produce a location. The list of operations is a follows:

WEIGHTING SCHEMES

HUBER Huber weighting (EQ 5.2-4).

BISQ Bisquared weighting (EQ 5.2-5).

SIN Andrew's sin weighting (EQ 5.2-6).

JEFF Jeffrey's weighting (EQ 5.2-7).

L2 least squares (L_2 norm).

L1 least absolute value (L_1 norm).

LOCATING

AOL Perform consistency check (Section 5.1) and use the arrival order location (Section 5.3.1) as the current location. Fix depth to 5 km.

STA1 Use the station with the earliest arrival as the current location. Fix depth to 5 km.

LOC3 Three parameter location: SEEK a new location using the specified weighting scheme while allowing origin time, latitude and longitude to vary with depth fixed at its current value.

LOC4 Four parameter location: from the current location, SEEK a new location using the specified weighting scheme while allowing origin time, latitude, longitude, and depth to vary.

The data set of hand picked arrival times were used to compare the results of five different location methods: BI (AOL, BISQ, LOC3, LOC4), [1] HUBER (AOL, HUBER, LOC3, LOC4), L1 (AOL, L1, LOC3, LOC4), L2AOL (AOL, L2, LOC3, LOC4), and L2 (STA1, L2, LOC3, LOC4). Since it was expected that JEFF and SIN would produce locations similar to BI they were not studied extensively. [2] The major differences in location were usually in the final depth as one might expect. Generally, the BI and L2 locations were the most different with the HUBER and L1 location either in between or close to one of the other locations. BI gave more reasonable depths and epicenters for events which are normally hard to locate such as surface events or events outside the array. Since the accuracy of the arrivals is usually good, the BI and L2 epicenters were often in relative agreement with each other. However, in 25 percent of the events, the epicenters differed by more than 4 km. and the depths differed by more than 6.5 km.

Most of the time, a robust method such as BI performed better than L2. <u>However, the performance is crucially dependent on the starting location used</u>. For example, for one quarry blast L2 gave a surface location whereas BI gave a depth of 6 km. In this case, arrivals from stations close to the epicenter were down weighted since the initial depth was 5 km and thus these arrivals were earlier than expected. Further iterations with the depth fixed increased the residuals for

---

[1] I.e., the method called BI consists of applying AOL followed by BISQ followed by LOC3 followed by LOC4. Doing LOC3 before LOC4 tends to reduce the number of iterations and increase the region of convergence (Buland, 1976; see below). For both LOC3 and LOC4 the velocity model used was a 6.25 km./sec. medium.

[2] See Chapter 3 for an analysis of the residuals using the BI locations.

these stations more and thus they were further down weighted. Thus eventually, all of the arrivals containing depth information were removed by the robust weighting!

Allowing the depth to vary starting from the trial location leads to a proper depth. However, it is a good idea to fix the depth for the first few iterations since this increases the region of convergence (Buland, 1976). Convergence is usually only a problem if the event is outside or on the edge of the array. Thus the following algorithm is recommended: Determine a trial location using AOL, BISQ, LOC3. If this location is outside the array, return this location; if not, return the location determined by AOL, BISQ, LOC4.

5.3.3. ROBUST VERSUS LEAST SQUARES LOCATIONS

In this section, the results of a study of the effect of arrival time errors on the BI and L2 locations for one event, 116.958, are given. The location and residuals for the L2 location of this event are shown in Exhibit 5.6, and the distribution of stations is shown in Exhibit 5.7.

This event was chosen for several reasons. Both BI and L2 give the same location and the residuals are small. Also, the event is typical of the smaller events recorded by CEDAR. [1] Since there are only seven stations reporting arrivals the effect of arrival time errors should be quite pronounced; both BI and L2 should perform worse than normal. The station geometry also has typical problems. The diameter of the reporting network, 50 km., is quite small and the distribution of the stations is not ideal. Station wml is 3.5 km from the epicenter and an error there should strongly affect the depth of the event. Station sgl is the most distant station and it is the only station that is significantly south of the epicenter. Thus the location should also be strongly affected by error at this station.

To investigate the effect of arrival time errors on the location, errors of -8, -4, -2, -1, 0, 1, 2, 4, and 8 seconds were added to a station and the event was relocated using BI and L2. This was repeated three times using stations wml, sgl, and nwr as the contaminated station.

---

[1] Roughly a third of the events are reported by less than 10 stations, a third are reported by 10 to 19 stations, and a third with 20 or more stations.

Exhibit 5.8 a shows the path of the L2 epicenters as a function of error at each of these stations. Generally, negative errors push the epicenter toward the station and positive errors push it away from the station. The location diverges (has extremely large latitude or depth) for larger positive errors at stations wml and sgl.

Exhibit 5.8 b show the paths for the corresponding BI epicenters. In most cases, the epicenters remain within a few kilometers of the actual epicenter. The effect of error at nwr appears to be almost completely removed whereas, only the +2 second error at wml provides a bad location. The effect of positive error at sgl is removed but the effect of negative error appears to be about as bad as that of L2. For comparison, Exhibit 5.9 shows the effect of error using L1. Except for the case of positive errors at nwr and wml, L1 is not significantly more robust than L2.

It is helpful to plot how errors at nwr, wml, and sgl affect the residuals at the other stations as shown in Exhibits 5.10, 5.11, and 5.12. Exhibit 5.10-a shows that for L2 locations, as the magnitude of the error at nwr increases, the residuals at the other stations are also increased. This inflation of the residuals is a general property of L2 and it is also shown in Exhibits 5.11-a and 5.12-a. Exhibit 5.10-b shows that BI essentially eliminates the residual inflation and the arrival time at nwr is always identified as an outlier. It is interesting that in the process of identifying the +1 and +2 second errors at nwr the weighting at wml is reduced somewhat, thus increasing its residual slightly.

Exhibit 5.11 b shows that once again BI generally correctly identifies WML as the outlier. It does make as mistake when wml has an error of +2 seconds. In this case, wlk, the second nearest station, is considered the outlier.

Exhibit 5.12 b shows that SGL is correctly identified as an outlier only when its errors are positive. For negative errors, sup is considered the outlier although the other residuals do remain somewhat inflated indicating that the location is poor. The reason that BI does so poorly in this case is that errors at sgl provide a strong influence on the location. Since sgl is quite separated from the other stations arrival time errors at sgl have a stronger leverage on the final solution. Also, its position biases the arrival order locations to be closer to sgl. As pointed out in Section 5.3.1, the perpendicular bisectors between pairs of stations are concentrated toward the center of the array. The position of sgl then biases the arrival order location to the south. Also, since there are only seven arrivals, the arrival order location is poorer than usual and the consitency sieve is not particularly effective.

In conclusion, it is interesting that a single event can provide so much useful insight into the robust location problem. Since it is an extreme case, it brings out the worst in both BI and L2. As expected L2 is strongly affected by arrival time errors. A single outlier can inflate the residuals at other stations. This can cause good arrivals to appear as outliers or moderate outliers to be hidden. Thus an L2 location should not be used as the initial location in a robust location scheme. Although the L_1 norm provides a good initial location for linear robust regression, it unfortunately does not appear to be significantly more robust than the L_2 norm for the nonlinear robust regression required for locating earthquakes.

Generally, BI appears to be quite capable of identifying outliers and providing robust locations. However, a good initial location is quite important; and in most cases, the arrival order location is good enough. However, one should never expect that only a handful of arrivals can provide a particularly robust location no matter what location method is used. As with the normal least square method (Buland, 1976) poor station geometry can be expected to take its toll on the robust location.

5.4. COMBINING THE PICKING AND LOCATION ALGORITHMS

Initially, 2 tests of the picking algorithm described in Chapter 4 were made with the trigger threshold, trigthr, set at four and five times the background noise amplitude, avabs. With trigthr set at five times avabs, a significant number of arrivals were missed; but although more arrivals were picked when trigthr was reduced to four times avabs, enough false picks were also made that the epicenters determined from them were worse than when the higher threshold was used.

Based on this, the following iterative picking and locating algorithm was used:

1. Put all available stations on a list to be picked, except for the following bad stations (Section 2.2): wwv, coa, cok, crr, hdg, ing, run, scy, slu, sup, adl, blu, coq, lhu, lcl, fma, dc5.

2. Deglitch stations bmt and twl before picking them (Section 4.1).

3 Pick stations with tp = 0.5; trigthr = 4.5; and codathr = 4.0.

4. Locate event using algorithm BI (Section 5.3.2).

5. Repick all stations for which no pick was recorded or if it was more than 0.6 seconds from its theoretical arrival time starting two seconds before its theoretical arrival time. The parameter SAM uses depends on the distance of the station from the event:

DISTANCE (KM.)	LLOMAX (SEC)	TRIGTHR (* AVABS)	CODATHR (* AVABS)
< 46	1	4	4
< 100	2	4	4
< 170	3	4	3
< 250	3	4	3
$\geq$ 250	NOT REPICKED		

6 Relocate event using new set of picks.

7 Display results.

The idea behind this algorithm was that on the first pass, the algorithm would pick up enough arrivals that a rough location could be determined. The second pass, armed with both a reduced threshold and a location should be able to clean up most of the bad picks from the first pass.

Since the velocity structure used could not predict the travel times to better than $\pm$ 0.3 seconds (Kanamori and Hadley, 1975), arrivals with residuals smaller than 0.6 seconds (to be on the safe side) were not repicked. Although SAM could have been restarted closer to the theoretical arrival time, starting two seconds before it removed any chance of the theoretical arrival time biasing the final pick. Only two iterations of picking and locating were performed since the picking algorithm usually would not have improved significantly on the third try especially since the advice it received (new picking parameter) was fairly weak. Basically, second pass pick parameters were adjusted with distance so that SAM would be more willing to accept weaker, more emergent arrivals than it did on the first pass. No amplitude or frequency requirements were put on the second pass picks since as yet they cannot be predicted very well. Although this algorithm is quite simple, its results are quite encouraging as is shown in the next section.

5.5. EXHIBITS

5.1. Algorithm SEEK finds the minimum of a loss function (see text.)

5.2. Example of the arrival order location method for a 5 station network. The numbered triangles indicate the station locations. The lines are the perpendicular bisectors for each pair of stations. The lines are labled to indicate the pair of stations they belong to. For example, the line which is near the upper left corner of the Exhibit is the bisector for stations 1 and 3 (see text).

5.3. This exhibit is similar to the previous one except that it is for a 20 station network. The 190 bisectors shown partition the network into an extremely large number of regions and thus the arrival order location should be very close to the epicenter.

5.4. The distance between the first station to record an arrival and the final epicenter is plotted versus the distance between the arrival order location and the final epicenter. The line shown has a slope of 1. Note that the arrival order location is nearer the final epicenter than the nearest station in all but four cases.

5.5. Scatter plot of the number of arrivals recorded for an event versus the distance between the arrival order location and the final epicenter. The plot shows that as long as there are at least 10 arrivals, the arrival order location should be within 3 km. of the epicenter.

5.6. L2 location parameters (and 95% confidence limits) and arrivals for event 116.958.

5.7. Map showing the station distribution for event 116.958. The star indicates the epicenter.

5.8. The change in the epicenter due to error at one of three stations (nwr, wml, and sgl) for a) the L2 algorithm and b) the BI algorithm. The curves show how the epicenter varies as the amount of error at each station changes. Generally, negative errors push the epicenter toward the station and positive errors push it away from the station. (The plots are scaled in kilometers.)

5.9. The change in the L1 epicenter due to error at one either nwr, wml, or sgl. (The plot is scaled in kilometers.)

5.10. Residuals at other stations as a function of error at nwr for a) the L2 location and b) the BI location.

5.11. Residuals at other stations due to error at wml for a) the L2 location and b) the BI location.

5.12. Residuals at other stations due to error at sgl for a) the L2 location and b) the BI location.

EXHIBIT 5.1

ALGORITHM SEEK: FIND MINIMUM OF $P(\theta)$

$\theta = \theta_0$

repeat {

$\theta_p = \theta$

$d\theta=[A^TWA]^{-1}A^TW[y - f(\theta)]$

$d = 1$

while $(P(\theta+d\theta\cdot d) > P(\theta))$ {

$d = d/2$

if $(d < SMALL)$ return θ

}

$\theta = \theta + d\theta$

} until $((P(\theta_p) - P(\theta))/P(\theta) < EPS)$

return θ

end

<u>SYMBOL</u>	<u>MEANING</u>
θ_0	initial solution,
θ	final solution,
θ_p	previous solution,
$P(\theta)$	robust loss function,
y	vector of observed arrival times,
$f(\theta)$	vector of theoretical arrival times,
A	matrix of partial derivatives w.r.t. θ,
W	diagonal matrix of robust weights,
d	damping factor
SMALL, EPS	appropriately small numbers.

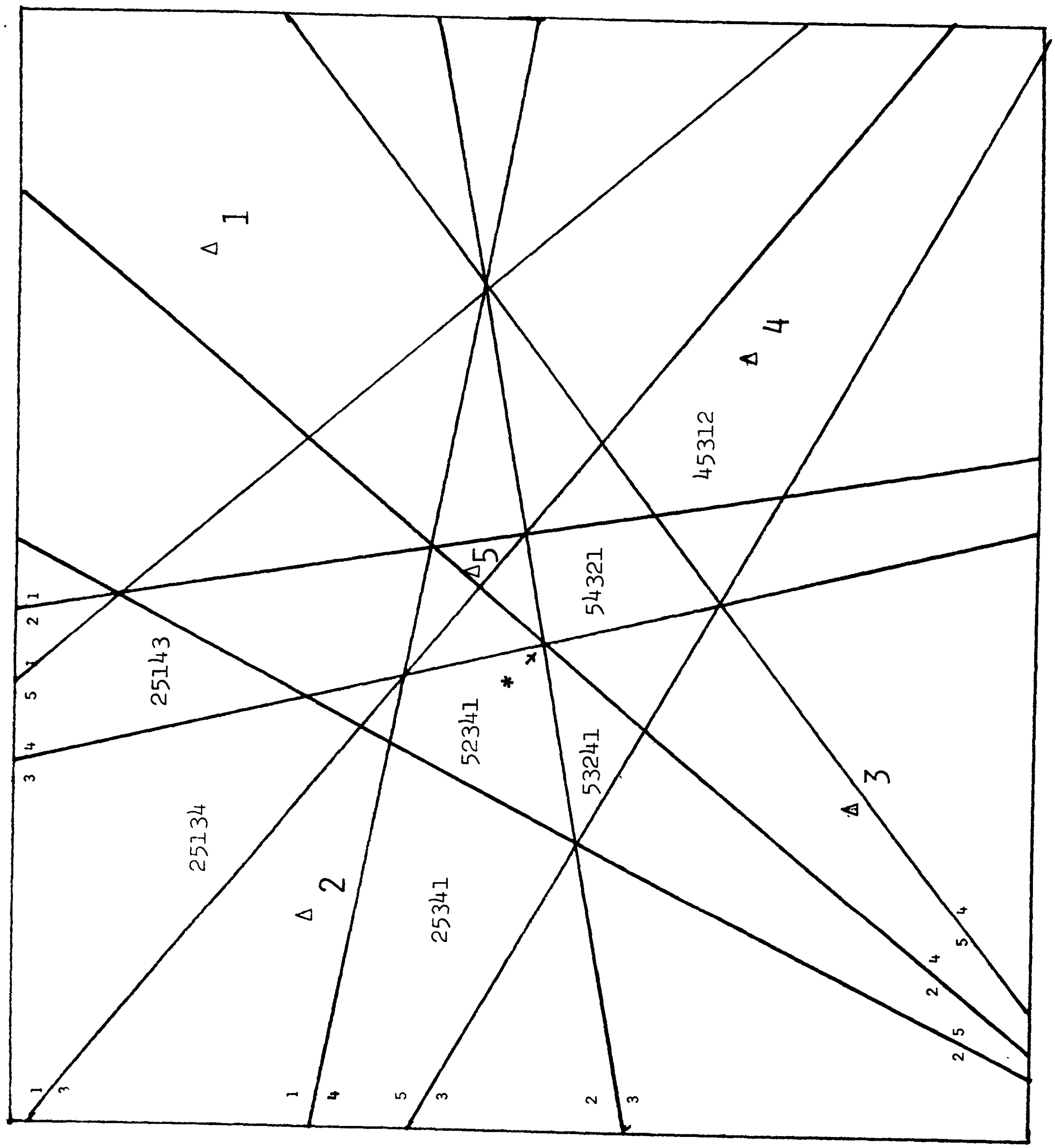


EXHIBIT 5.2

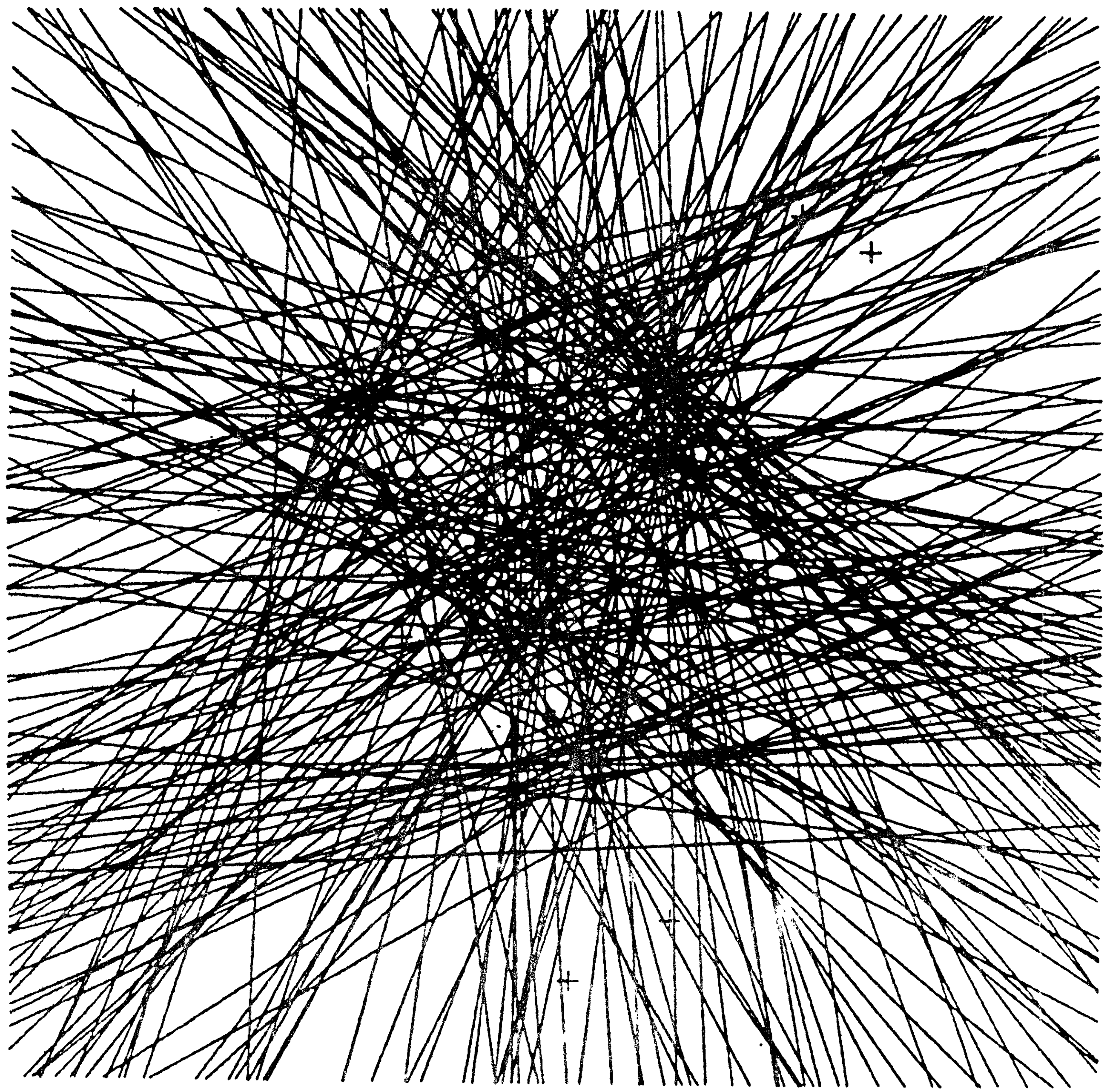


EXHIBIT 5.3

EXHIBIT 5.4

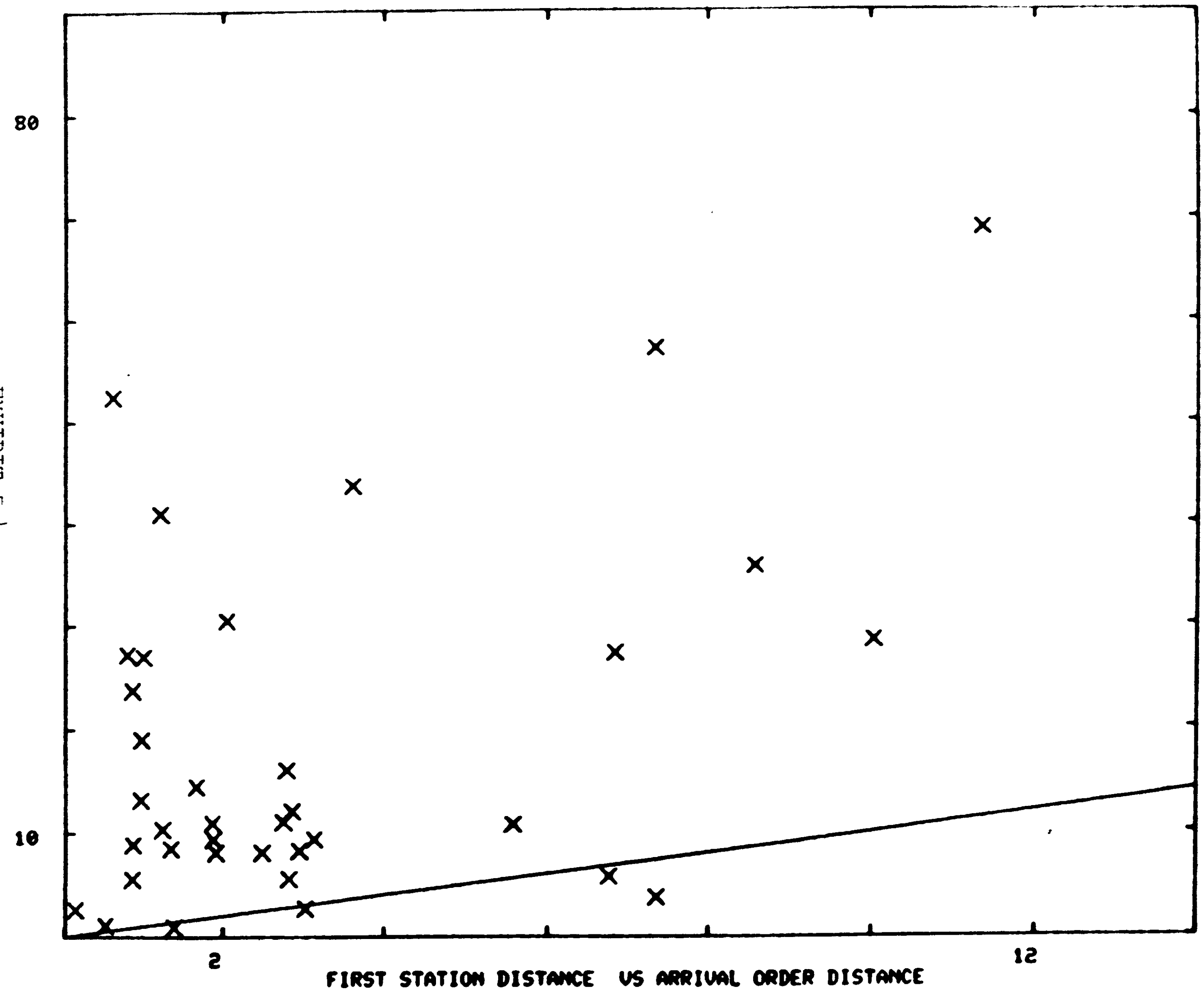


EXHIBIT 5.5

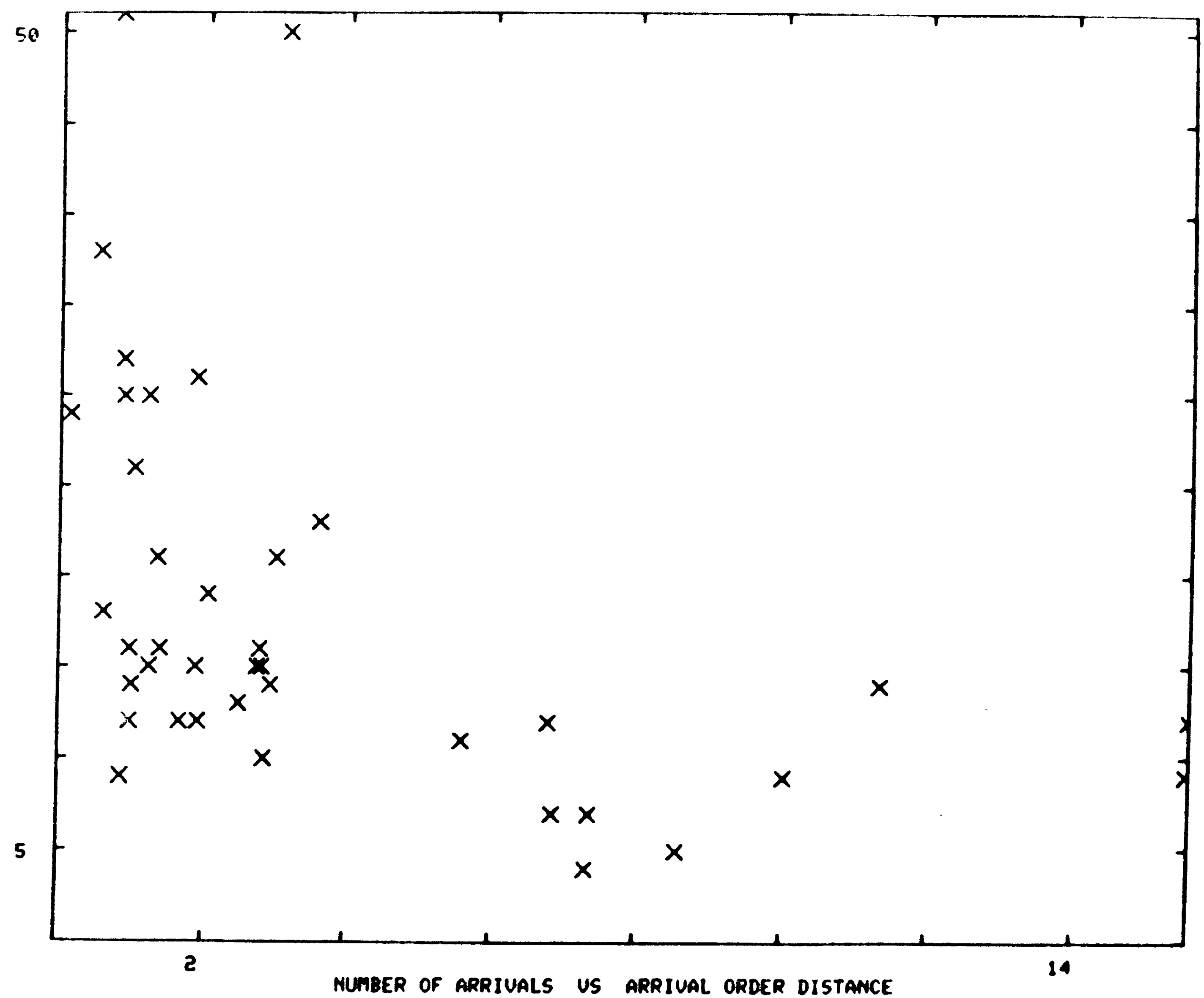


EXHIBIT 5.6
L2 LOCATION PARAMETERS FOR EVENT 116.958

ORIGIN TIME	LATITUDE	LONGITUDE	DEPTH
7.22±.10	33.006±.006	-115.587±.006	3.63±2.2

STATION	ARRIVAL TIME	RESIDUAL	QUALITY	DISTANCE	AZIMUTH
wml	8.03	0.01	100	3.50	286
wlk	8.90	-0.06	75	10.28	60
nwr	9.64	0.12	50	13.95	319
cli	9.78	-0.05	75	15.95	20
sup	10.78	-0.15	25	22.87	255
ing	11.48	0.08	75	25.91	94
sgl	13.90	-0.01	100	41.64	198

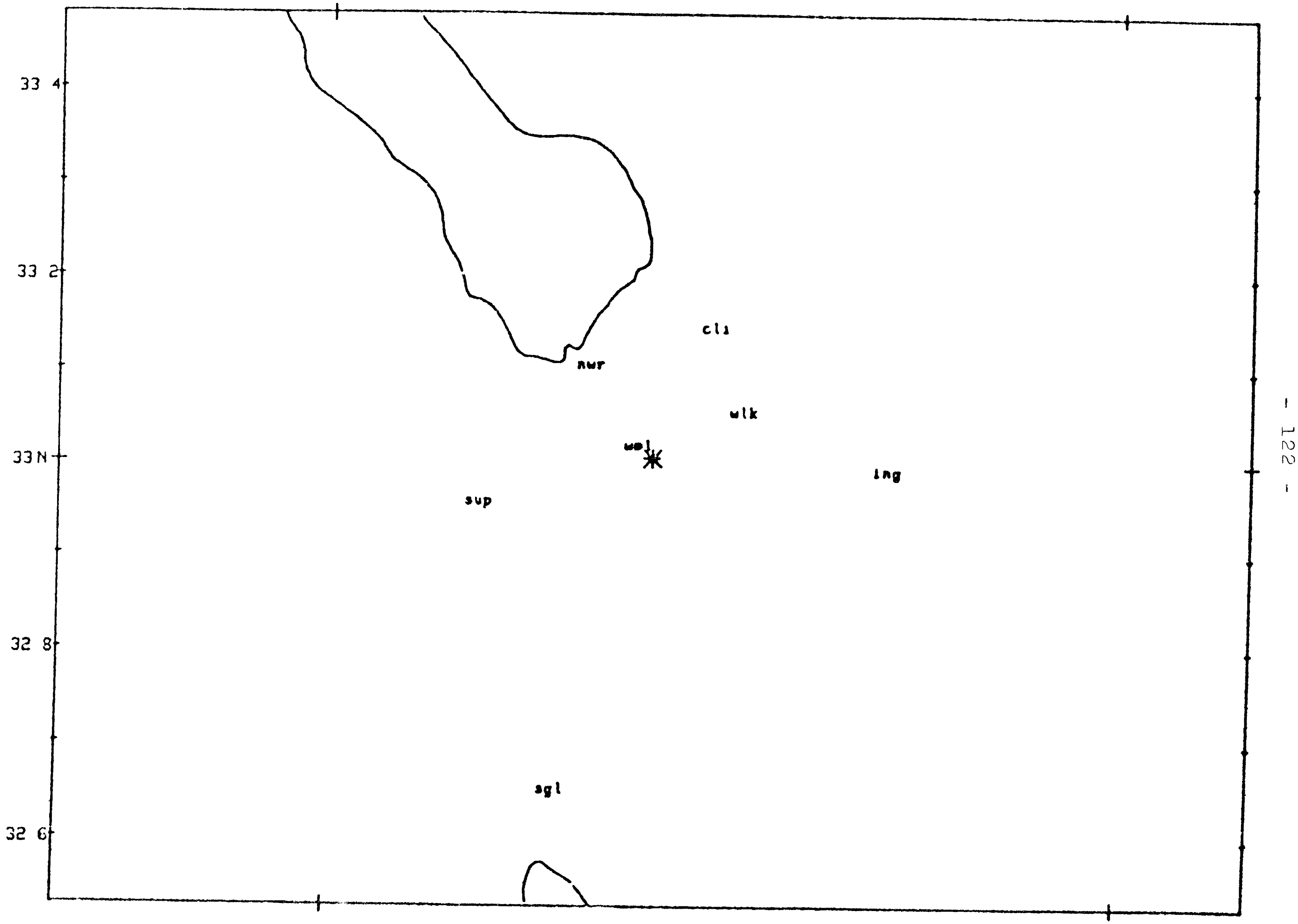
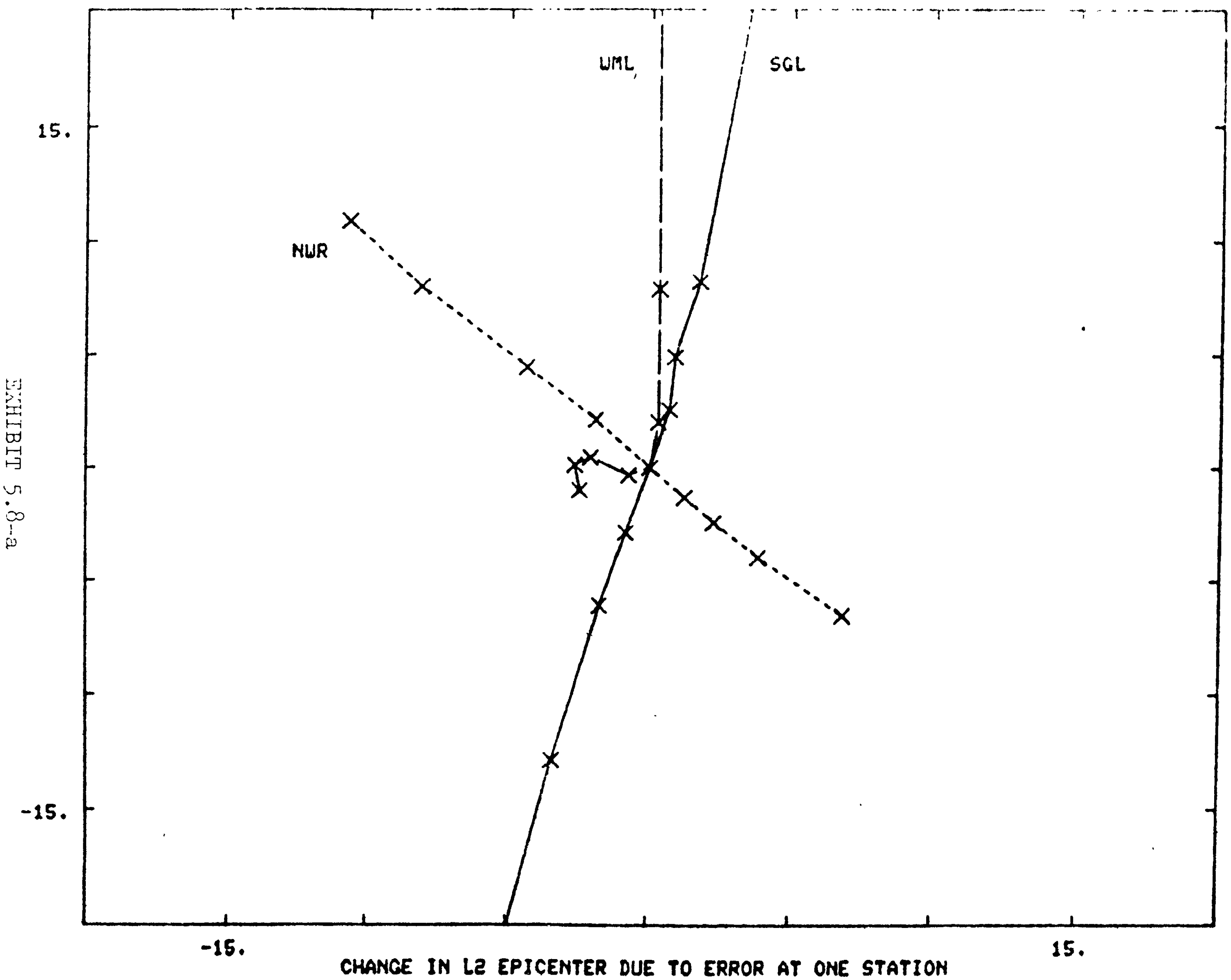


EXHIBIT 5.7

EXHIBIT 5.8-a



WML
SGL
NWR
15.
-15.
-15.
15.
CHANGE IN L2 EPICENTER DUE TO ERROR AT ONE STATION

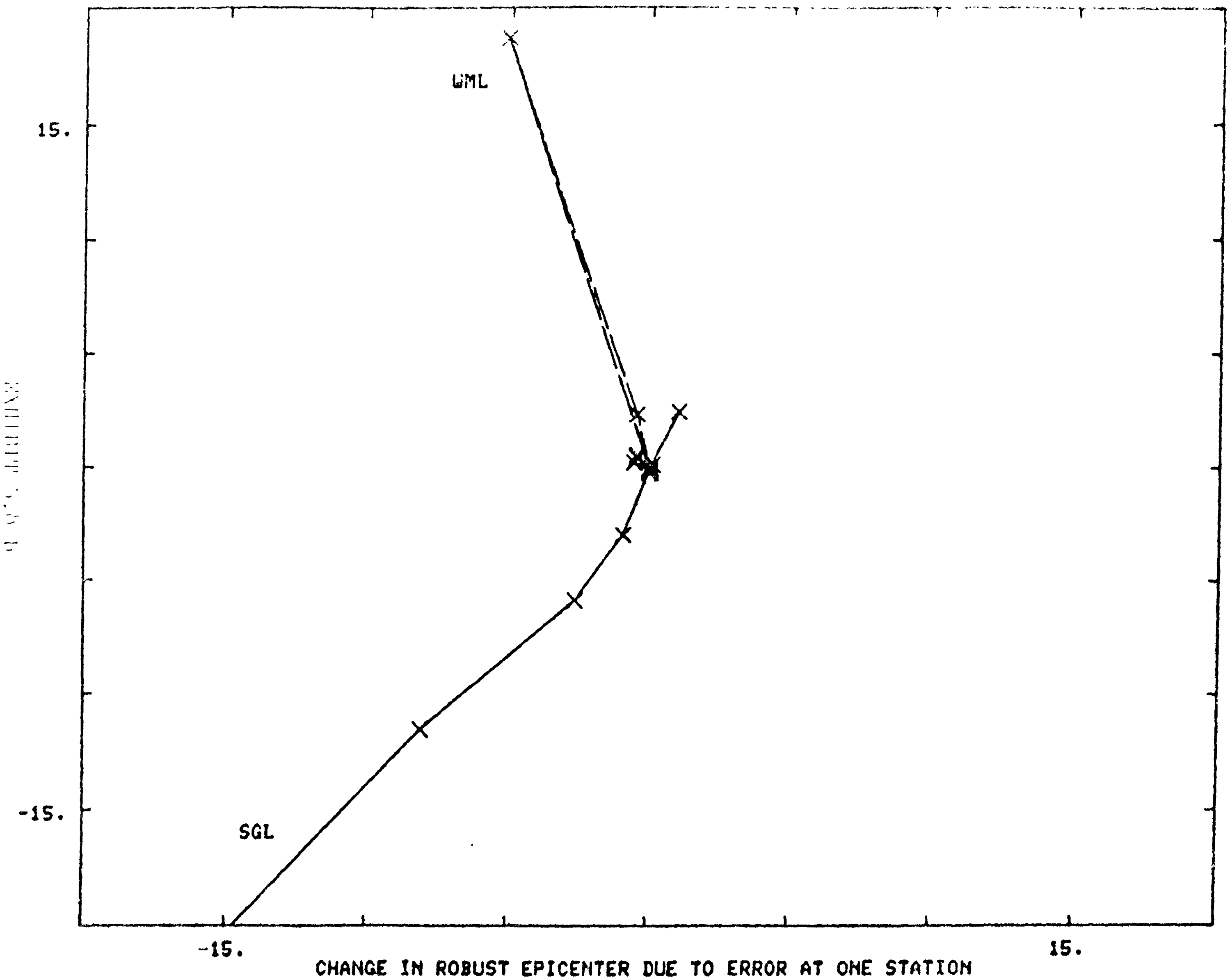

WML
SGL
15.
-15.
-15.
15.
CHANGE IN ROBUST EPICENTER DUE TO ERROR AT ONE STATION

EXHIBIT 5.9

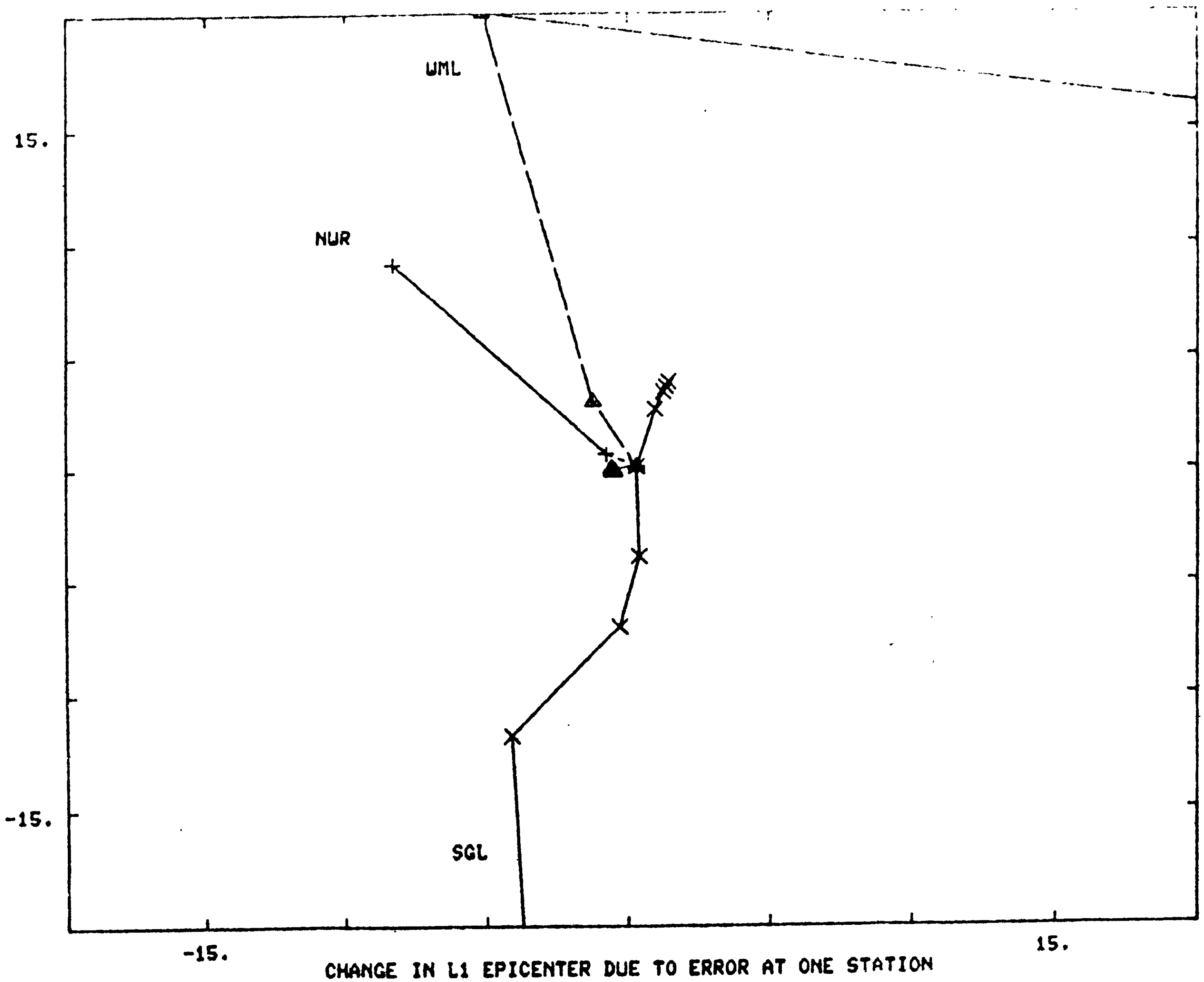
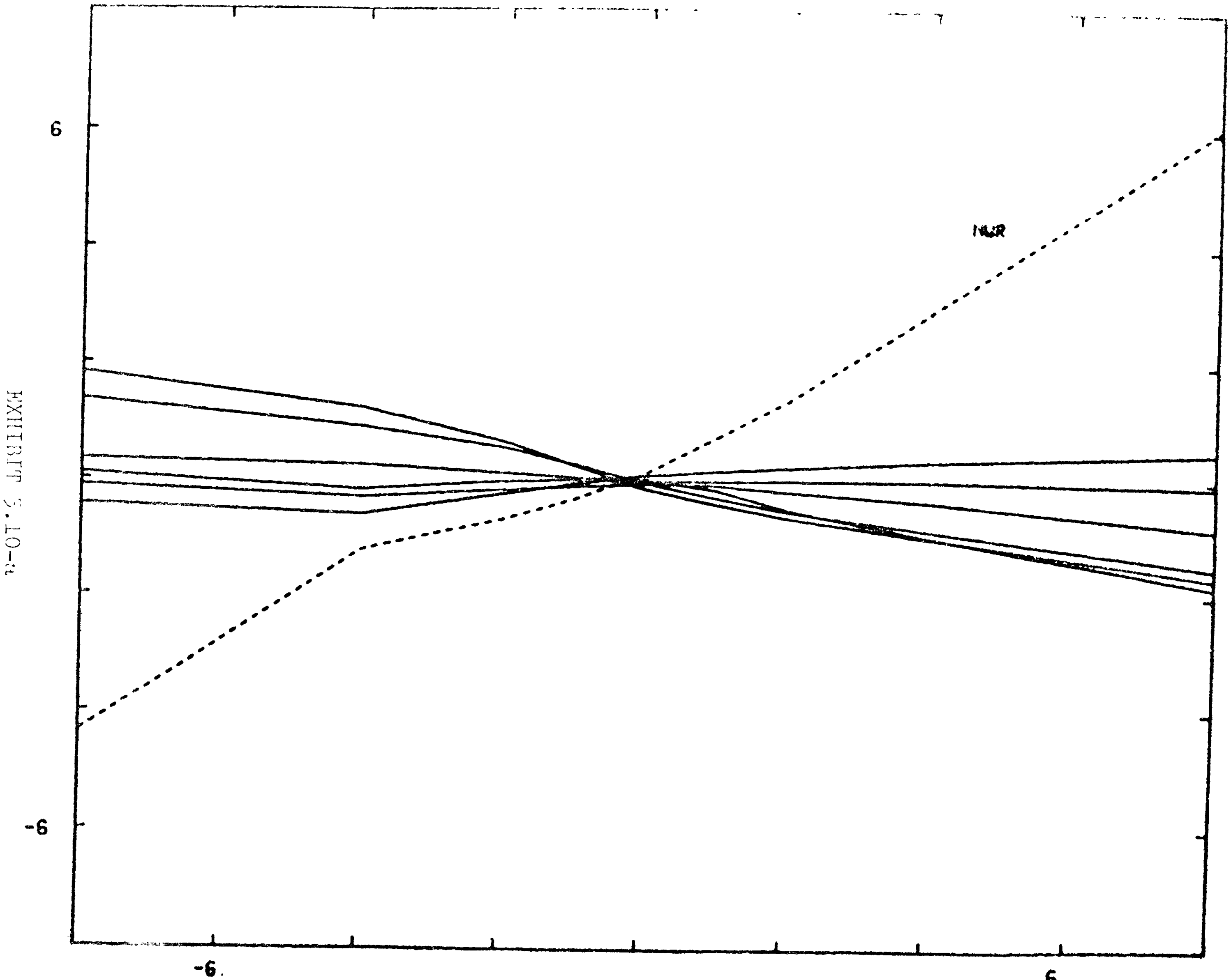
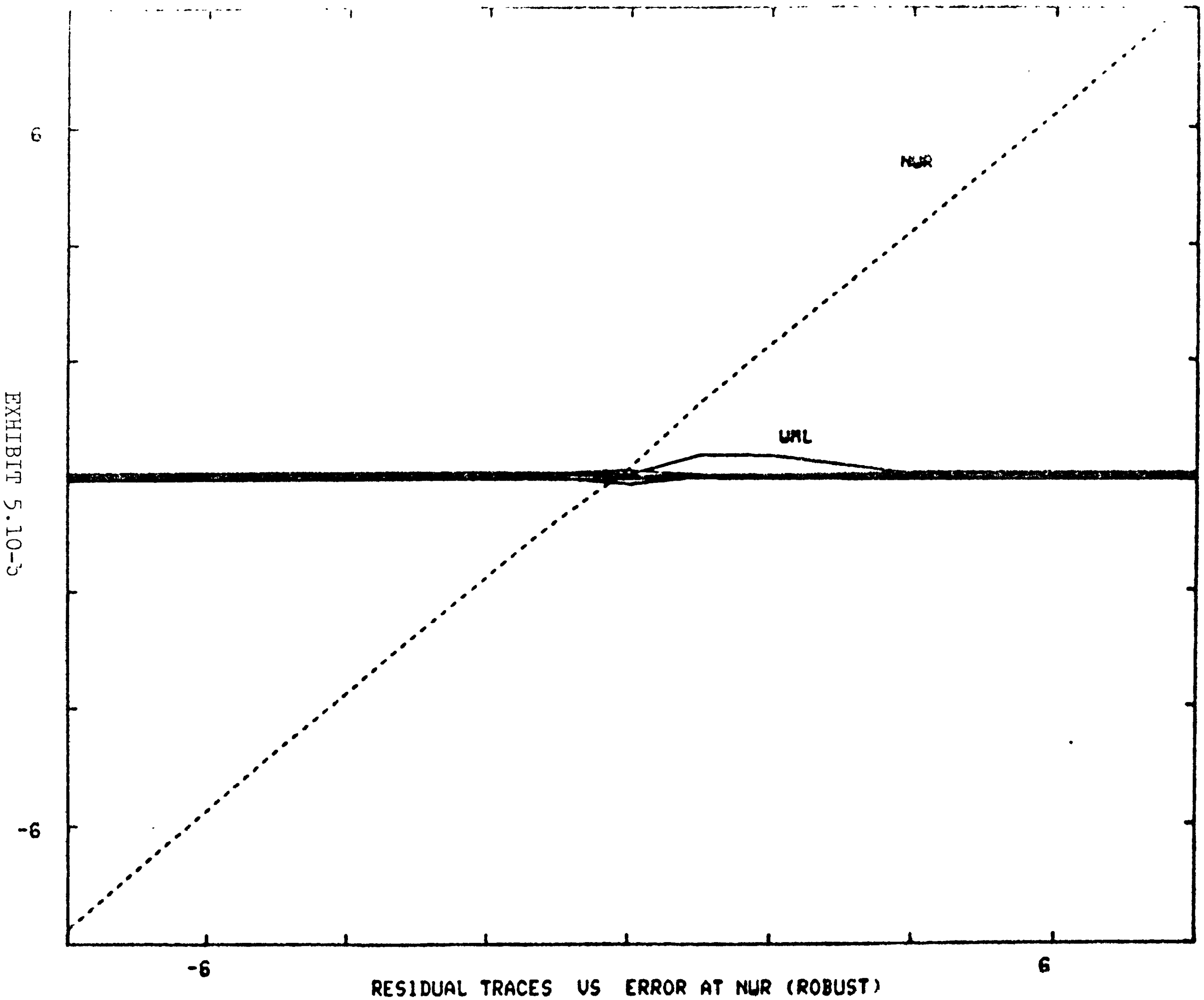

WML
NUR
SGL
15.
-15.
-15.
15.
CHANGE IN L1 EPICENTER DUE TO ERROR AT ONE STATION

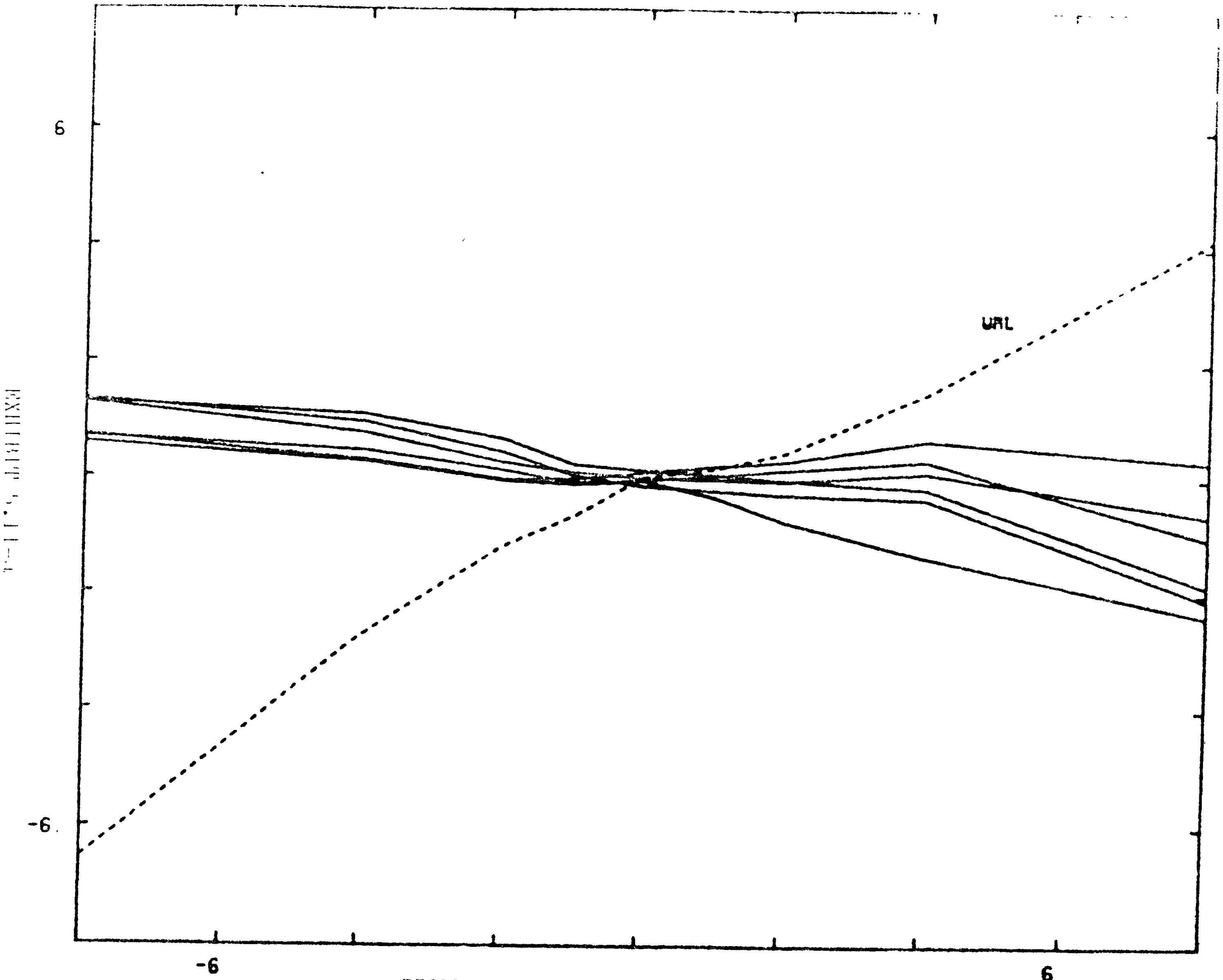
EXHIBIT 3.10-a



RESIDUAL TRACES VS ERROR AT NUR (L2)

EXHIBIT 5.10-3





RESIDUAL TRACES VS ERROR AT UML (L2)

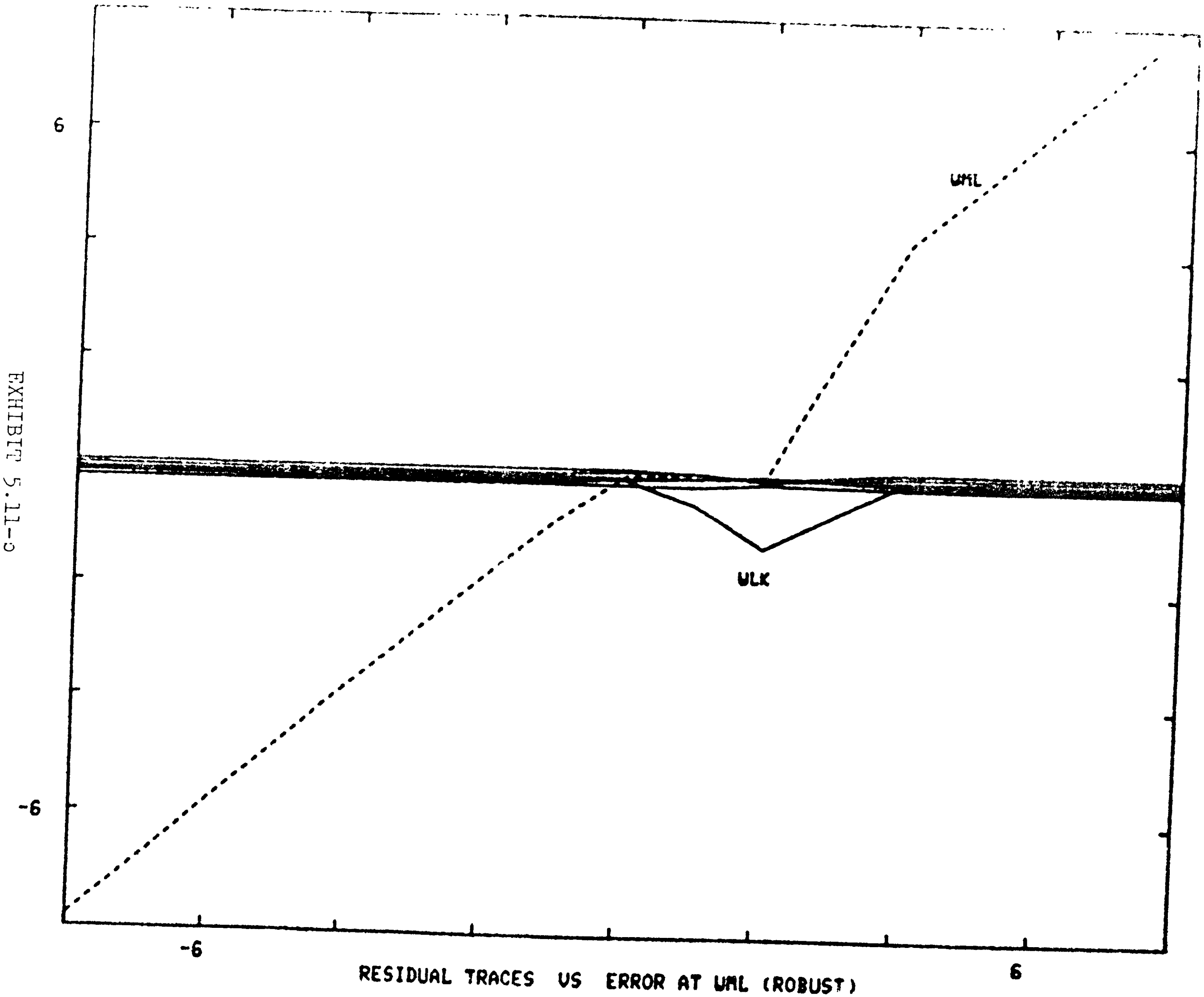
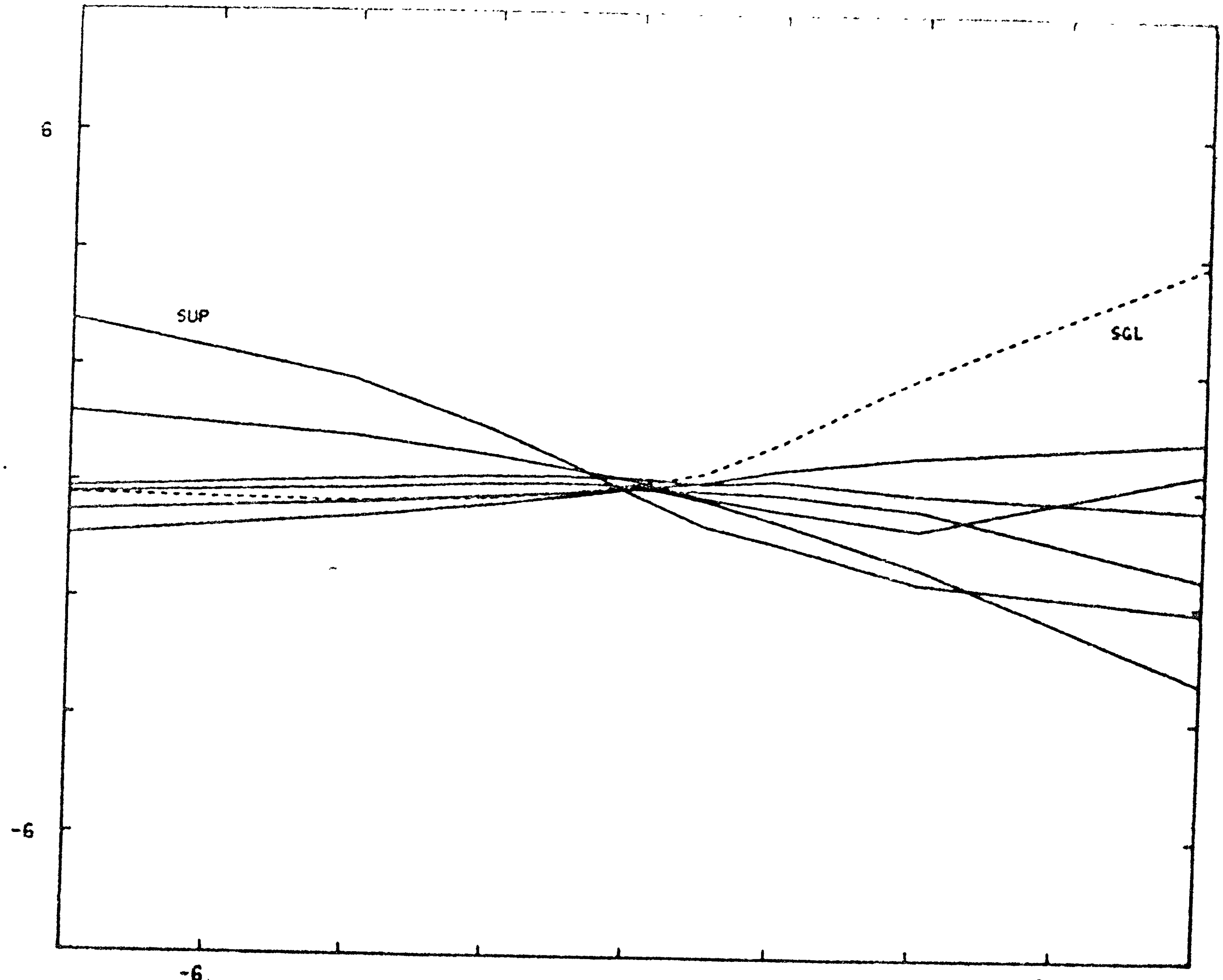


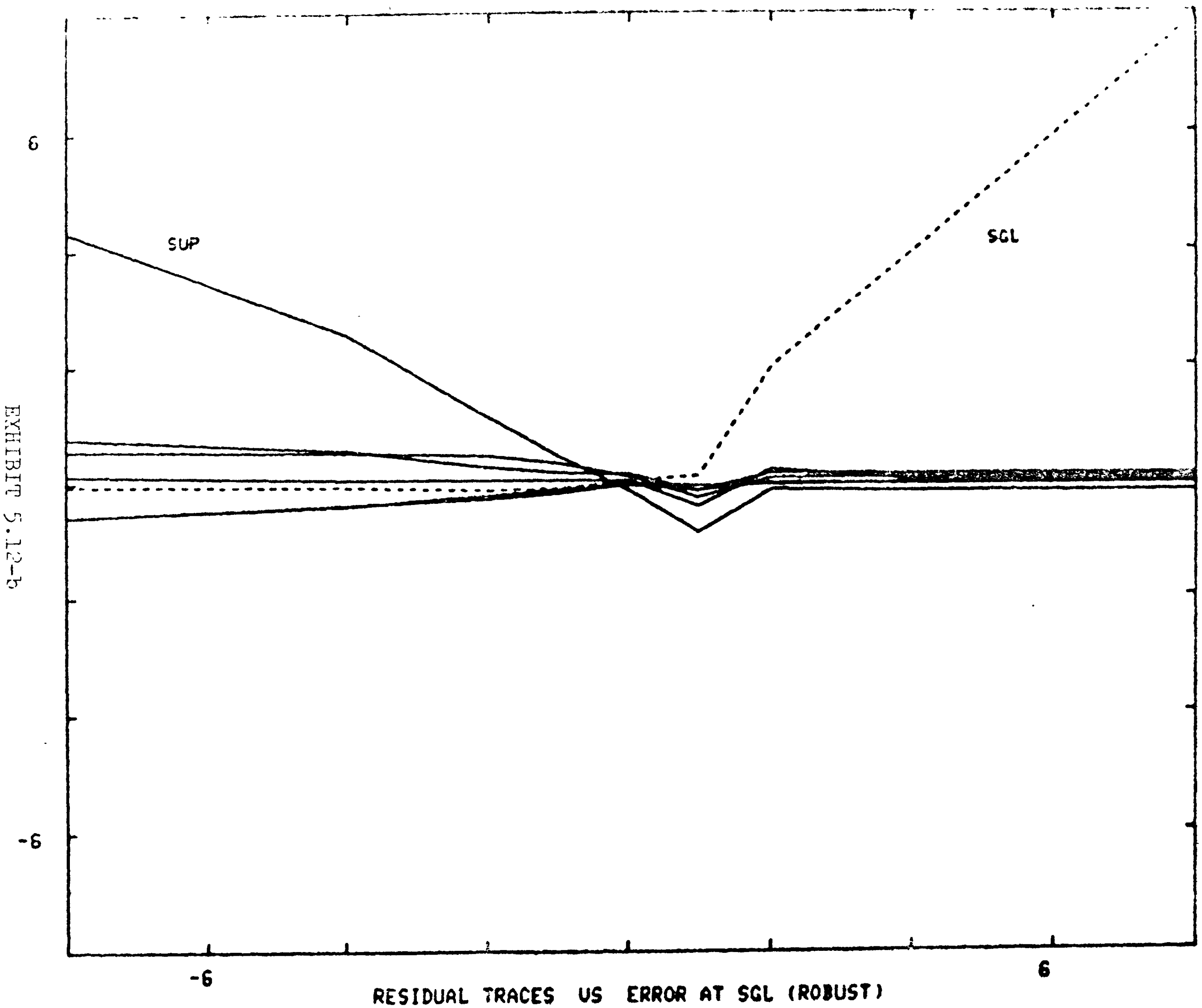
EXHIBIT 5.11-5

EXHIBIT 5.12-a



RESIDUAL TRACES VS ERROR AT SGL (L2)

EXHIBIT 5.12-b



6. COMPARISON OF HUMAN AND AUTOMATIC PROCESSING

Data from 25 typical events from the central portion of the network were used to compare machine and hand processing. The automatic processing scheme presented in the previous section was applied to 912 seismograms from these events. From these 912 seismograms, an analyst was willing to pick 407 P arrivals.

The final machine epicenters are shown in Exhibit 6.1, Two of the events (116.1720 and 117.242) had too few reliable arrivals to be located properly. Event 116.1720 is particularly interesting since even seismologists had difficulty with this as shown in Chapter 4. Except for these two events, the machine determined epicenters and depths agree quite closely to the hand determined ones. 75 percent of the epicenters were within 1.2 km. and 75 percent of the depths were within 1.9 km. of each other.

6.1. COMPARISON OF ARRIVAL TIMES

The dataset consisted of 407 hand picks. Histograms of the difference between the machine picked arrival time and the hand picked arrival time (M-H) for the different arrival qualities are shown in Exhibit 6.2. A median-hinge summary of M-H versus arrival quality is shown in Exhibit 6.3. For the Q100 and Q75 arrivals, 75 percent of the machine and hand picks agree to better than 0.04 seconds (two sample points apart). As the arrival quality decreases, the machine picks later than the seismologist. For Q50 arrivals, half the the machine picks are within 0.07 seconds of their corresponding hand picks and 75 percent are within 0.16 seconds. Since the dominant frequency for these weaker arrivals is around 8 Hz., the machine appears to be picking one or two half cycles later than the seismologist. For Q25 arrivals, this bias is even worse, and the algorithm is typically several half cycles late. Overall, 50 percent of the arrivals are within 0.04 seconds and 75 percent are within 0.12 seconds of the hand picked times.

A median-hinge summary of M-H as a function of distance is shown in Exhibit 6.4. The median M-H stays relatively small from 0 to about 110 km. Beyond this distance, the algorithm clearly tends to pick late. This is not unreasonable since Chapter 3. showed that hand picks also tended to be less reliable in this region due to a weakening first arrival and crossing branches of the travel time curve (See Exhibit 3.5.).

Another way to compare the machine and hand picks is to see how often the machine picked the same half cycle that the seismologist did. The picking algorithm provides the width of the half cycle it picked as the P arrival. Assuming the width of the first few half cycles are relativley constant, (M-H)/width gives an estimate of how many half cycles off the algorithm is.

Exhibit 6.5 shows the percent of arrivals for which the absolute value of M-H is less than n times the pulse width as a function of n for Q100, Q75 and Q50 arrivals. About 90 percent of Q100 and Q75 arrivals are less than one pulse width from their hand pick. Unfortunately, only about 55 percent of the Q50 arrivals are within one pulse width of their hand pick. Although, the weaker arrivals are more ambiguous and thus there is a a higher chance that the machine and hand picks would disagree the algorithm is too simple to pick these arrivals as early as a seismologist does. For weak arrivals, a seismologist uses frequency and wave shape information that the algorithm does not use.

Of the 407 hand picked arrivals, there were only 24 (7 percent) for which the machine did not return an arrival time. The number of missed picks for each arrival quality is given in Exhibit 6.6. The number of arrivals missed is quite small for the higher quality arrivals, and increases as the quality of the arrival decreases as expected.

It is of some interest to investigate why the six high quality arrivals were missed. Exhibit 6.7 a shows the two Q100 arrivals missed. The dotted line indicates the hand picked arrival time. In both cases, the picking algorithm was triggered by these arrivals but they were rejected because their coda was not long enough. This is not unreasonable since both of these arrivals have little energy and are fairly atypical. The four Q75 arrivals which were missed are shown in Exhibit 6.7 b. These arrivals are also quite weak, although in some cases, their onset is fairly clear. Again the picking algorithm is triggered by the arrivals but the arrivals are ultimately rejected.

Thus, the picking algorithm only misses the worst of the Q100 and Q75 arrivals. In fact, there is some indication that these arrivals should have received a lower quality estimate. A pick-summary for event 119.1311 is shown in Exhibit 6.8. Note that even though many of the arrivals are quite weak, every arrival is given the highest quality (Q100)!

There were 205 additional machine picks which had no corresponding hand picks. The travel time residuals, MRESID's, for these arrivals are shown in exhibit 6.9. These arrivals appear to come from three different sources; P arrivals, S arrivals, and false alarms.

There are 94 P arrivals which had residuals small enough to be considered in the location. For the most part, these are generally weak arrivals which were either missed or ignored by the seismologist. Although, these are not the best quality arrivals they are useful since they help the consistency check identify large errors and they provide the arrival order method with additional constraints. They are also important for a complete catalog of seismograms since many source-receiver paths may only have weak arrivals on them. <u>Thus through dogged determination, the algorithm was able to identify 25 percent more arrivals than the seismologist did</u>.

Sixty seven of the machine picks appear to be S arrivals and their residuals match the predicted S-P time as shown by the line in Exhibit 6.9. For all of these arrivals, the corresponding P arrival was very weak to non-existent since they were missed by the algorithm when very lenient parameter settings were used on the second pass. Clearly the second pass picking algorithm should have ignored any pick outside the time window in which a first arrival was expected. <u>The location algorithm identified these arrivals as S arrivals by their large positive (late) residuals and they were not used in the location</u>.

The times for the remaining 44 arrivals do not correspond to either P or S so they are considered false alarms. Since a low trigger threshold is used on the second pass, a false trigger in front of the actual arrival is not unlikely, and since the p-acceptor and coda-acceptor does find an event, the false trigger is considered an arrival. A third application of the picking-locating loop should have eliminated these early arrivals. Actually these arrivals are not true false alarms since the algorithm correctly identifies them as errors (by giving them zero weight). Thus an actual false alarm is an arrival which looks like a seismic arrival and occurs inside the the expected time window but is not of seismic origin. <u>Thus the true false alarm rate of the algorithm is near zero</u>.

As one might expect, it was usually only the weak arrivals which needed to be repicked. Of the roughly 600 seismograms for which an arrival was returned by the algorithm, 245 were repicked on the second pass. The percent of the arrivals which needed to be repicked for each arrival quality is shown in Exhibit 6.10. Only about 10 percent of the Q100 and Q75 arrivals were bad enough to be repicked, whereas 20 percent of the Q50 arrivals and almost half of the Q25 arrivals had to be repicked.

6.2. COMPARISON OF QUALITY ESTIMATE

It is difficult to compare the arrival quality assigned by hand and the standard error, TERR, assigned by the computer. The hand assigned quality is based on a five point scale which as pointed out above may not have been used consistently throughout the data set. It is also not clear whether the quality scale is a somewhat relative scale which may change from event to event. On the other hand, the computer assigned standard error is an absolute scale based on the noise level and the slope of the onset.

Actually, TERR should be more useful than the hand assigned quality since it can provide a confidence interval for the arrival. If we assume that the hand picked arrival times are always correct, M-H is an estimate of how far off the machine pick is from the actual onset. Then a plot of the cumulative percent of arrivals with the absolute value of M-H less than n time TERR can provide empirical confidence limits for the arrival times. Such a plot for Q100, Q75 , Q50 and all arrivals is shown in Exhibit 6.11. Only those machine picks that agree to within one pulse width were used since for picks with larger M-H TERR is not meaningful. About 85 percent of the Q100 and Q75 arrival times and 70 percent of the Q50 arrivals are within one TERR of their corresponding hand picked times. Thus TERR tends to over estimate the accuracy of the Q50 arrivals somewhat. Overall however, as long as the algorithm has picked the correct half cycle, one and two times TERR can be used for rough 80 and 90 percent confidence limits for arrival times.

6.3. EXHIBITS

6.1. Locations of the 25 events used to compare human and automatic processing.

6.2. Histograms of the difference between the machine picked arrival time and the hand picked arrival time, M-H (sec.), for each hand picked arrival quality.

6.3. A median-hinge summary of M-H versus arrival quality (See Exhibit 3.5).

6.4. A median-hinge summary of M-H versus distance (km.).

6.5. The percent of arrivals for which the absolute value of M-H is less than n times the pulse width versus n for Q100, Q75, and Q50 arrivals. This shows that for 90% of the higher quality arrivals the computer picks the same half cycle as the analyst did (see text).

6.6. The number of hand picked arrivals which the algorithm did not pick for each arrival quality.

6.7. A plot of the Q100 and Q75 hand picked arrivals which the algorithm missed. Note that these are all extremely weak arrivals.

6.8. Machine picked arrivals for event 119.1311. Each seismogram is annotated as described in Exhibit 1.1. Note that although many of the arrivals are quite weak, every arrival was given the highest quality (Q100) by an analyst.

6.9. Scatter plot of the travel time residuals of machine picks which had no corresponding hand picks as a function of distance. Most of the arrivals are near the predicted P or S (solid line) arrival time.

6.10. For each quality, the percentage of arrivals which were repicked.

6.11. The cumulative percentage of arrivals with the absolute value of M-H less than n time the computer assigned standard error, TERR, for those arrivals for which the machine and hand pick were less than 1 pulse width apart. This shows that one and two times TERR can be used as rough 80 and 90 percent confidence limits for the arrival times (see text).

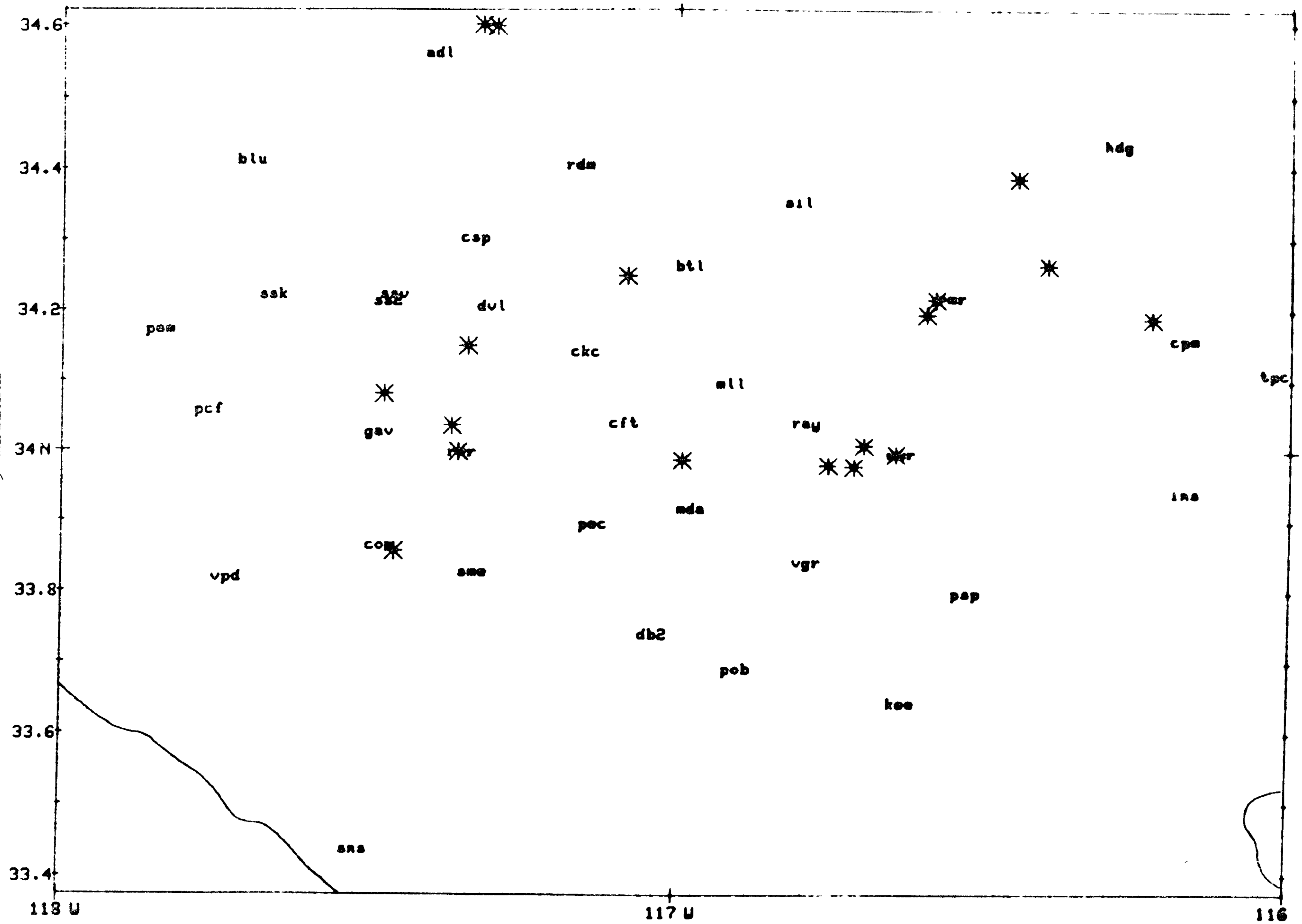

34.6
34.4
34.2
34N
33.8
33.6
33.4
118 W
117 W
116
adl
blu
rdm
hdg
sil
csp
btl
ssk
dvl
pem
ckc
cpm
tpc
mll
pcf
cft
ray
gav
ins
mda
pec
vgr
vpd
sme
pap
db2
pob
kee
sns

EXHIBIT 6.1

EXHIBIT 6.2-a

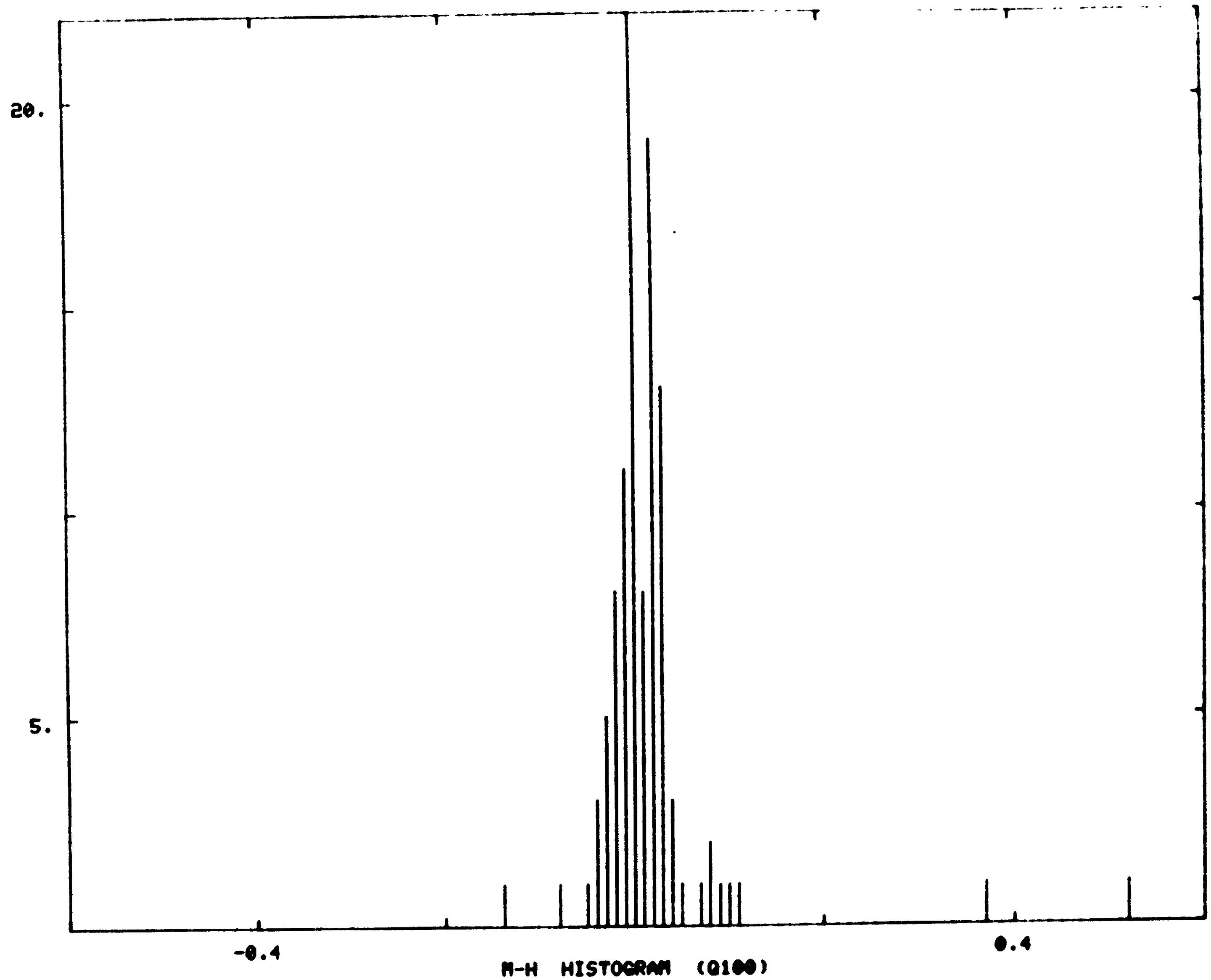


EXHIBIT 6.2-b

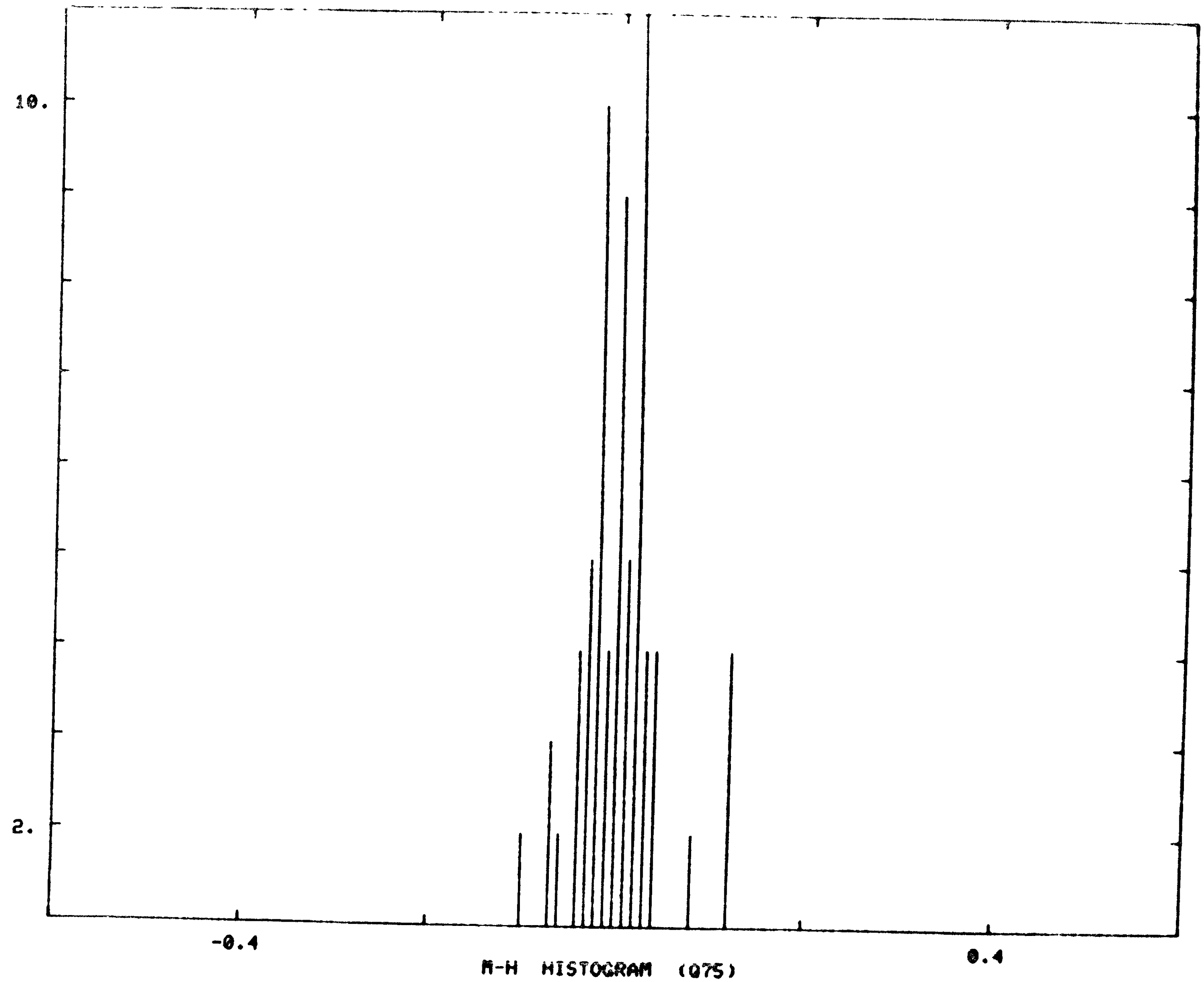
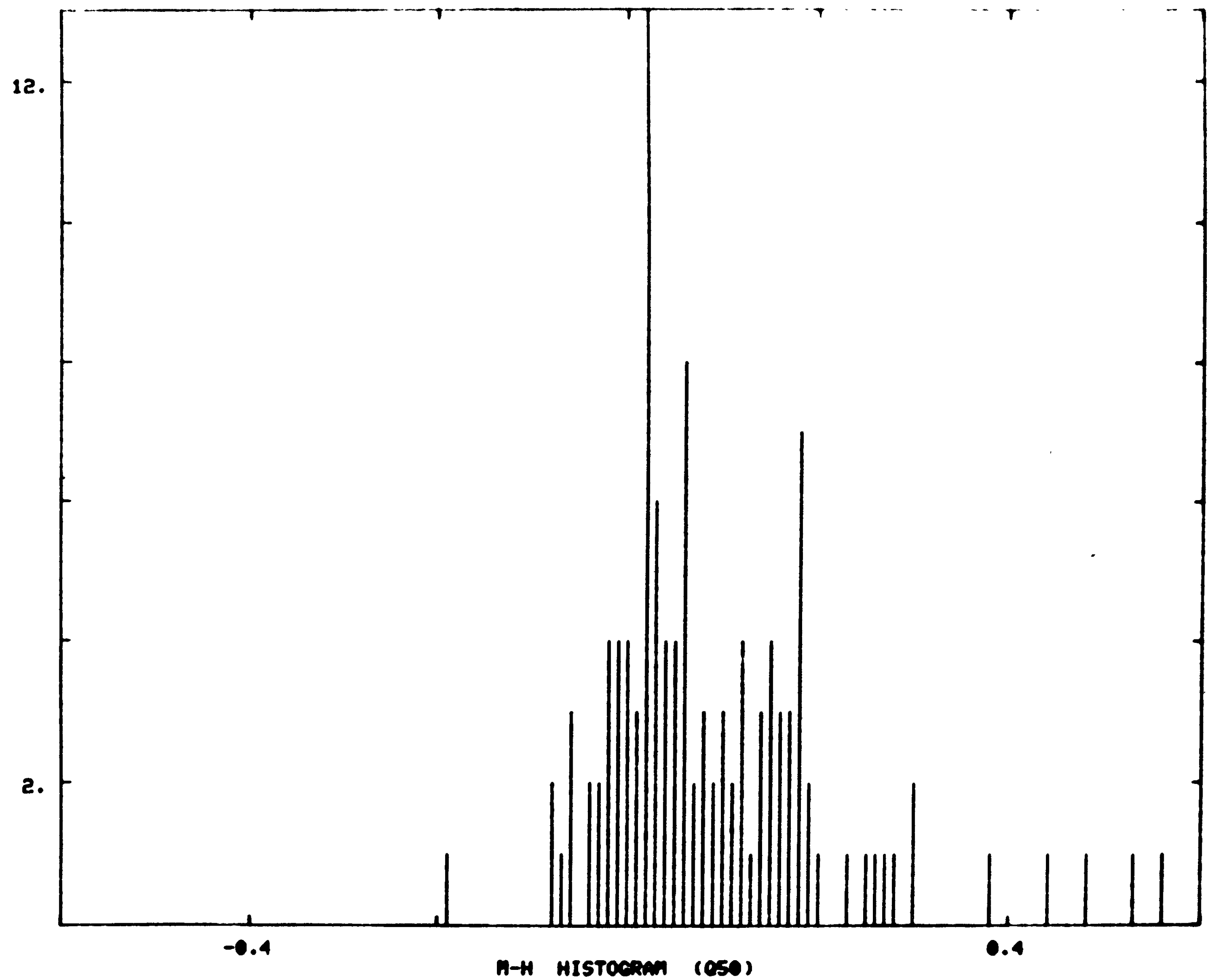
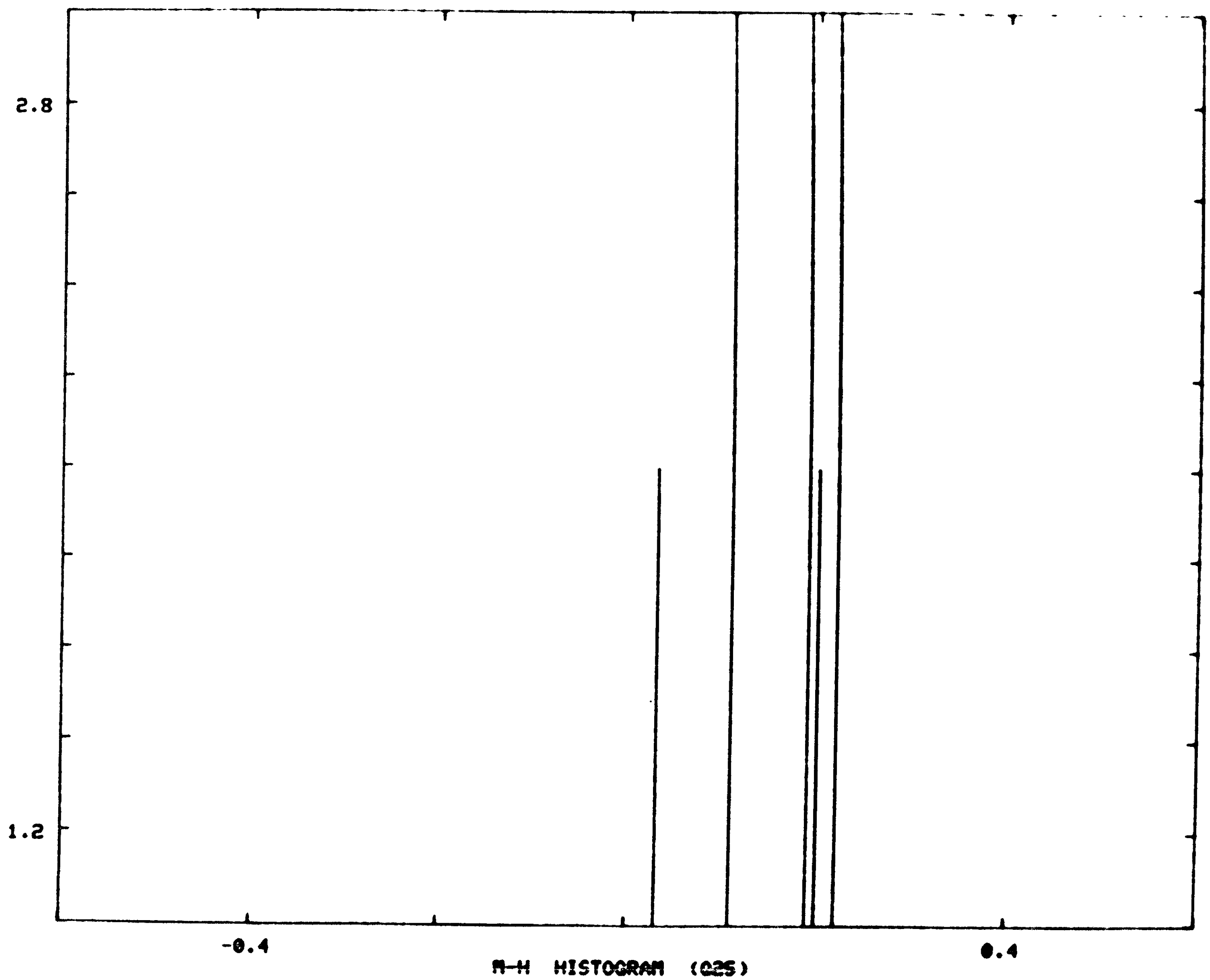


EXHIBIT 6.2-c




2.8
1.2
-0.4
0.4
M-H HISTOGRAM (Q25)

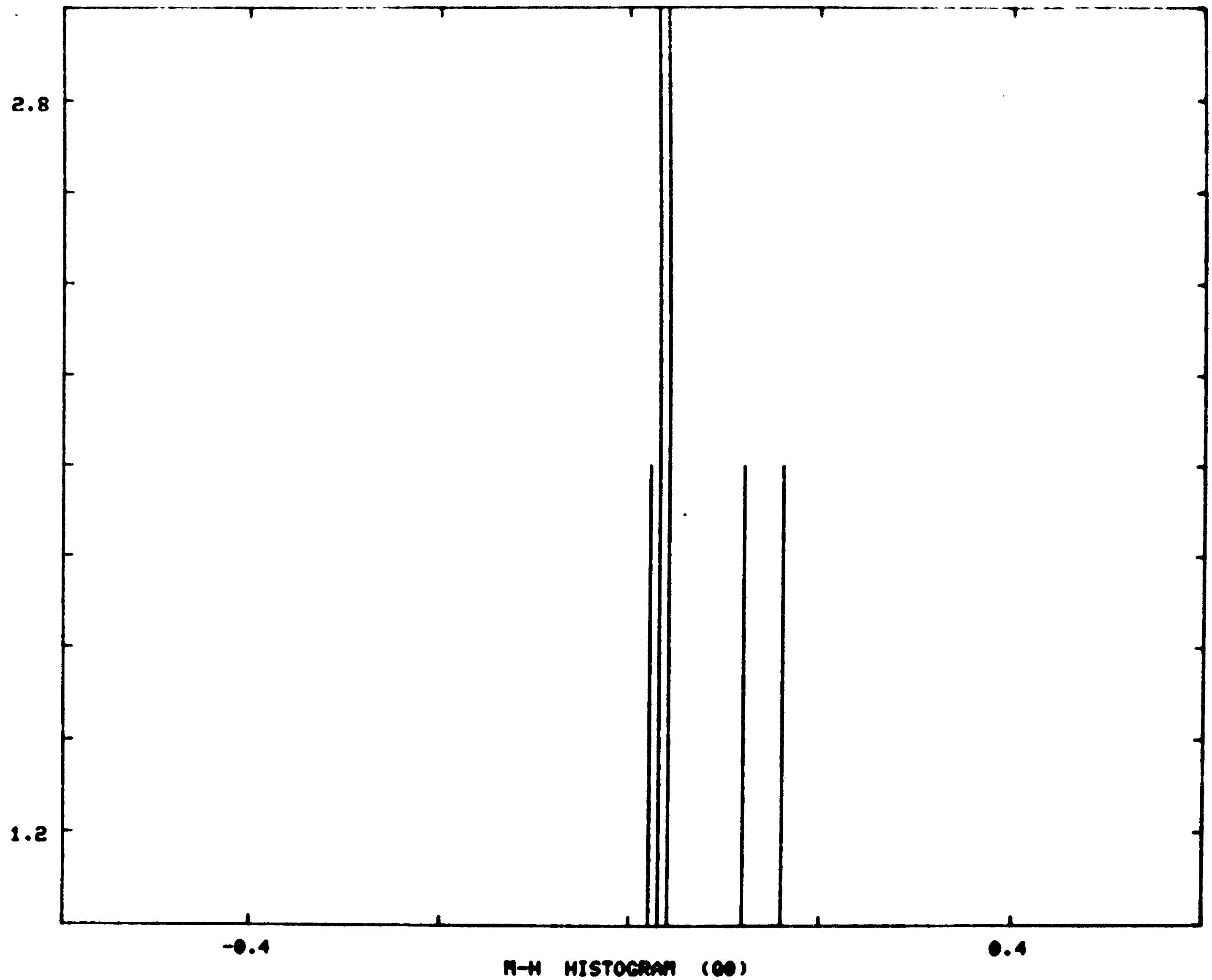


EXHIBIT 6.2-e

EXHIBIT 6.3

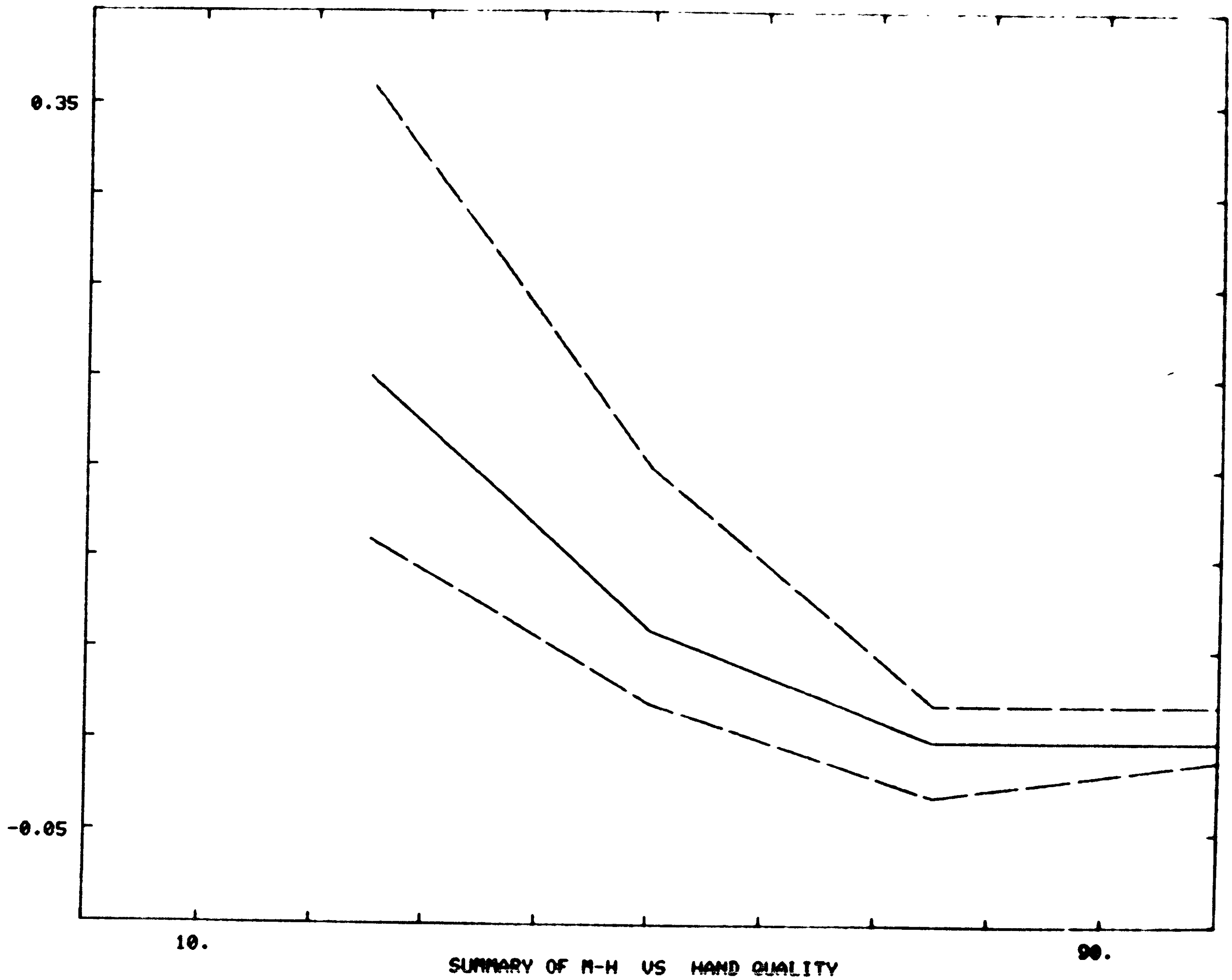
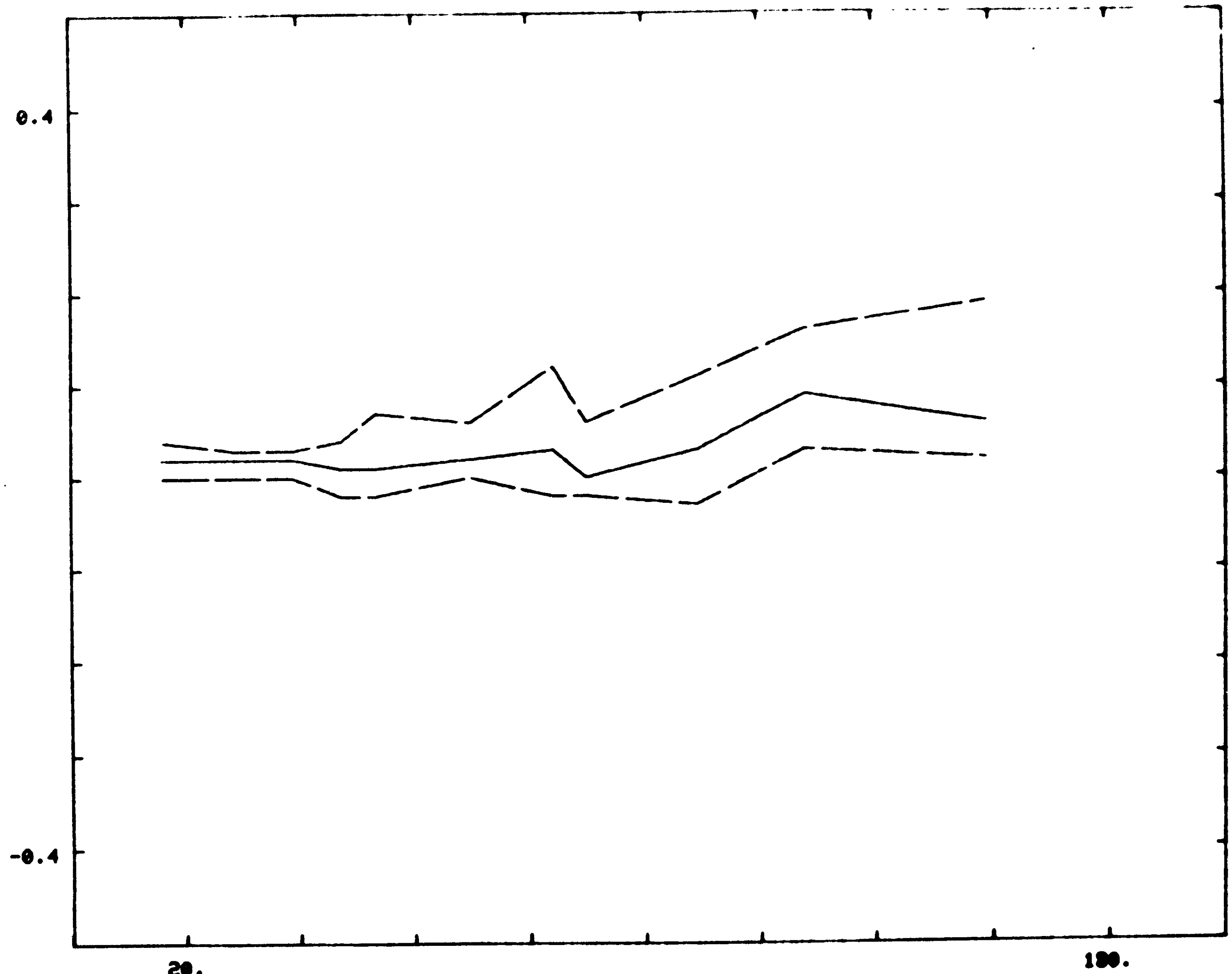


EXHIBIT 6.4



SUMMARY OF M-H VS DISTANCE (KM.)

EXHIBIT 6.5

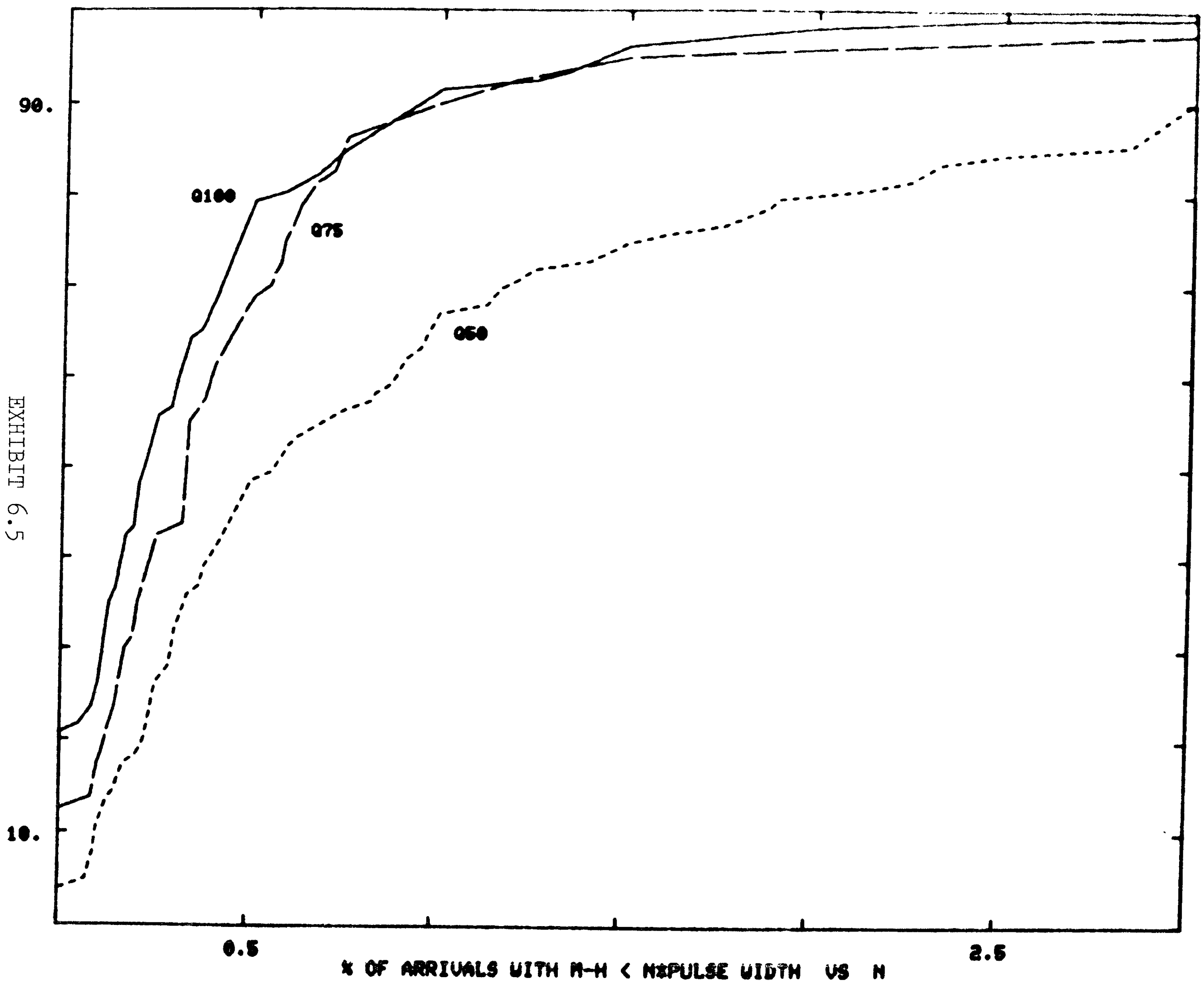


EXHIBIT 6.6

NUMBER OF MISSED PICKS FOR EACH QUALITY

QUALITY	MISSED	TOTAL	% MISSED
100	2	104	2
75	4	76	5
50	5	115	4
25	13	38	34

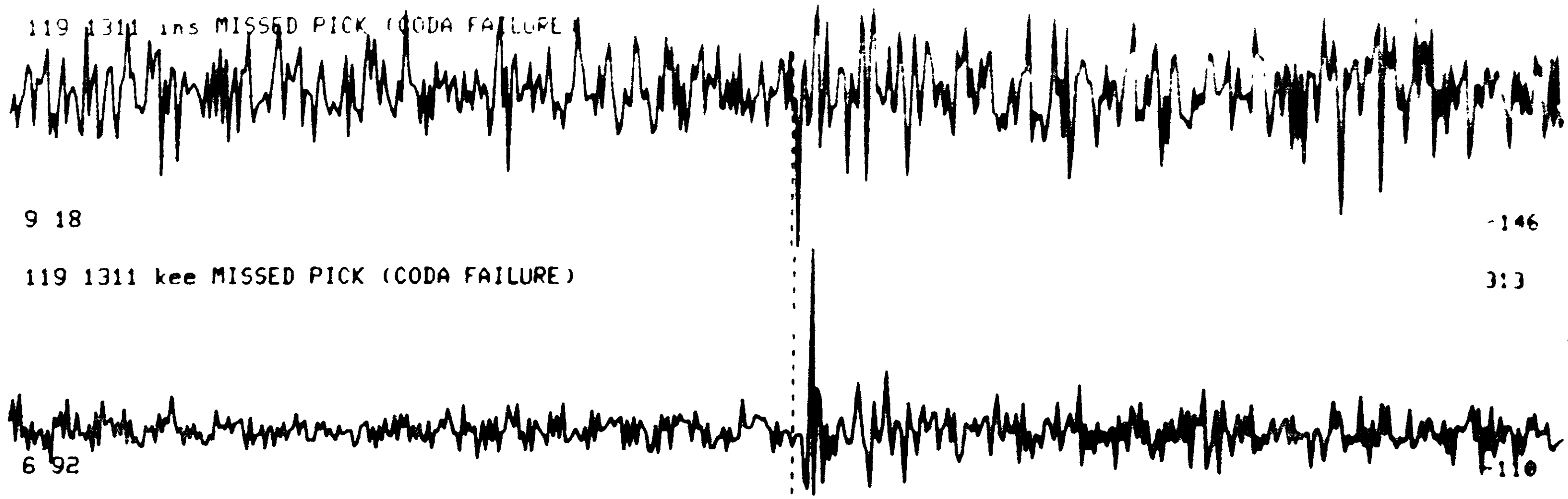

119 1311 ins MISSED PICK (CODA FAILURE)
9 18
-146
119 1311 kee MISSED PICK (CODA FAILURE)
313
6 92
-110

EXHIBIT 6.7-a

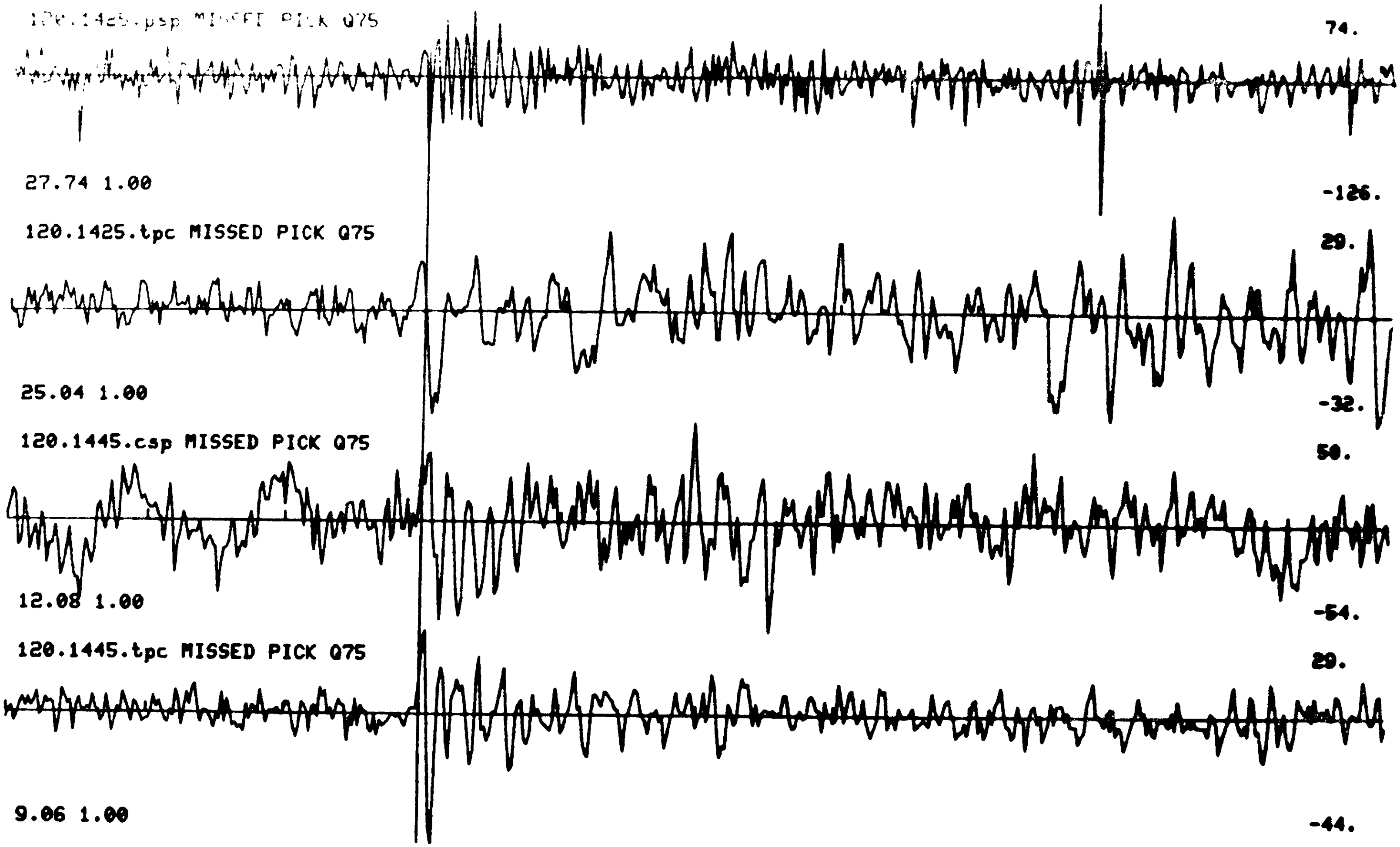

120.1425.psp MISSED PICK Q75
74.
27.74 1.00
-126.
120.1425.tpc MISSED PICK Q75
29.
25.04 1.00
-32.
120.1445.csp MISSED PICK Q75
50.
12.08 1.00
-54.
120.1445.tpc MISSED PICK Q75
29.
9.06 1.00
-44.

EXHIBIT 6.7-b

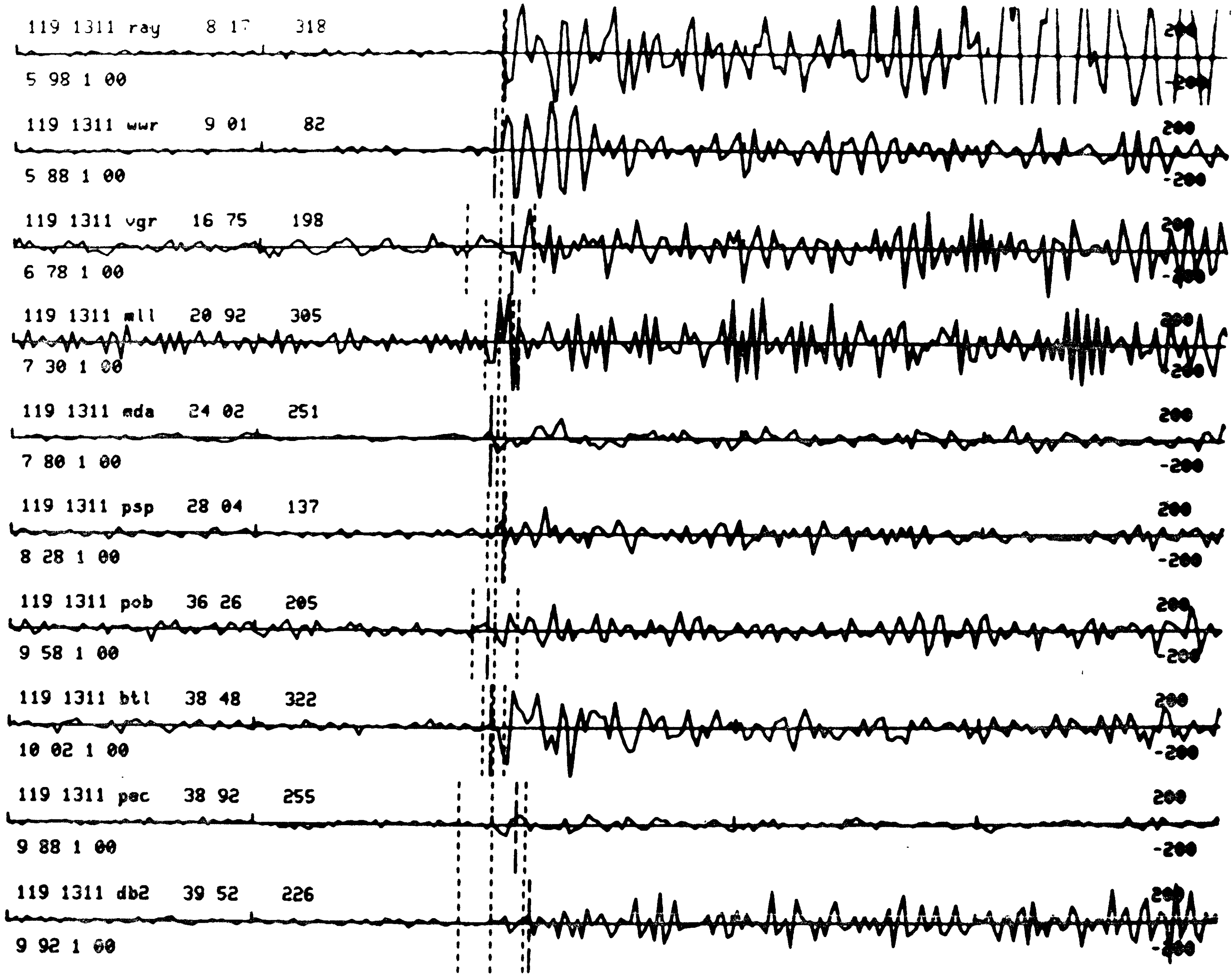

119 1311 ray 8 17 318
5 98 1 00
200
-200
119 1311 wwr 9 01 82
5 88 1 00
200
-200
119 1311 vgr 16 75 198
6 78 1 00
200
-200
119 1311 mll 20 92 305
7 30 1 00
200
-200
119 1311 mda 24 02 251
7 80 1 00
200
-200
119 1311 psp 28 04 137
8 28 1 00
200
-200
119 1311 pob 36 26 205
9 58 1 00
200
-200
119 1311 btl 38 48 322
10 02 1 00
200
-200
119 1311 pec 38 92 255
9 88 1 00
200
-200
119 1311 db2 39 52 226
9 92 1 00
200
-200

EXHIBIT 6.8

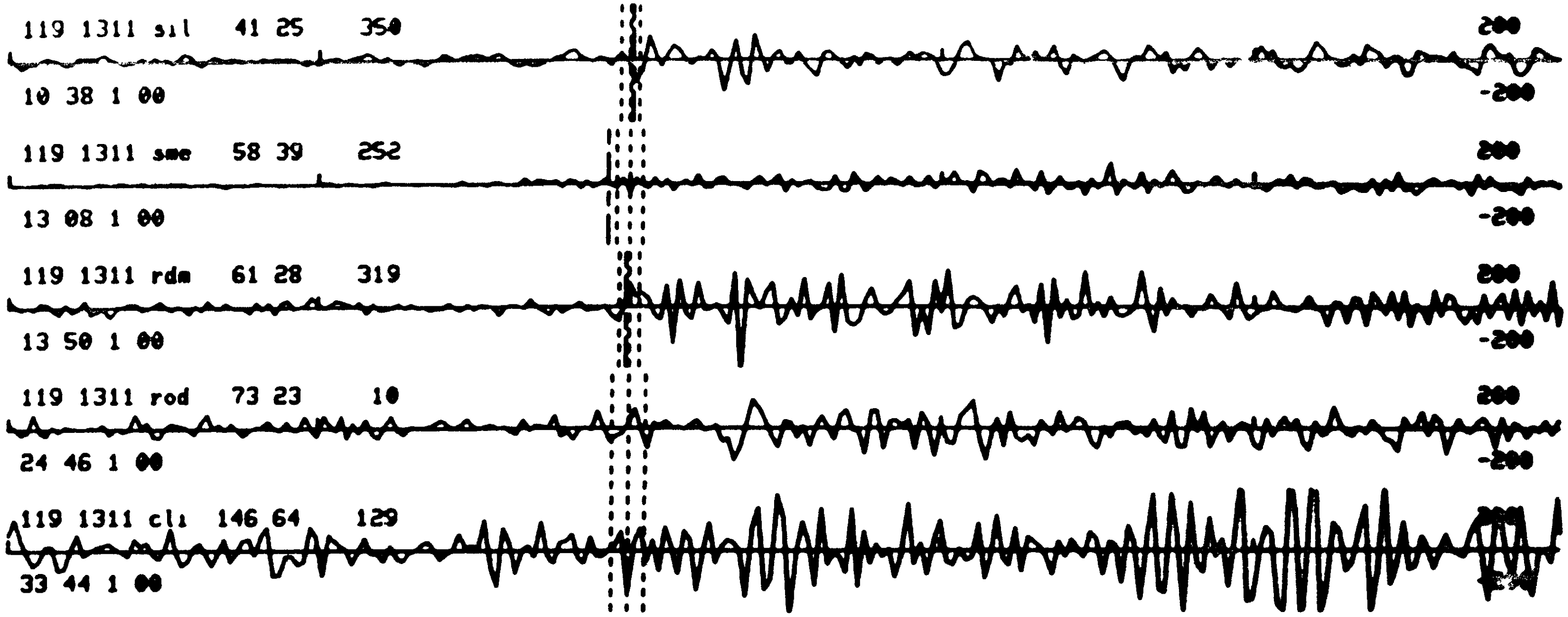


EXHIBIT 6.8 CONTINUED

EXHIBIT 6.9

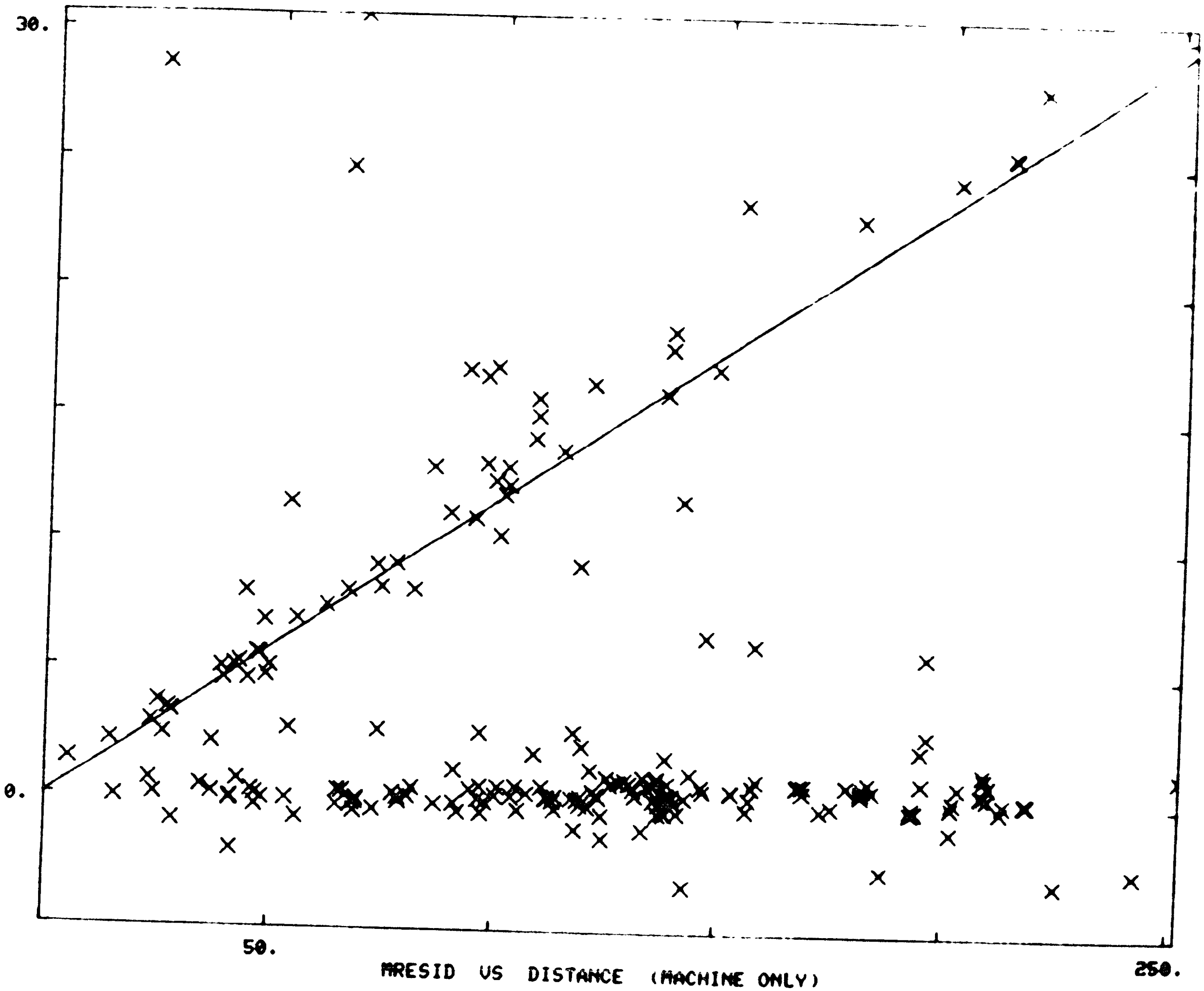
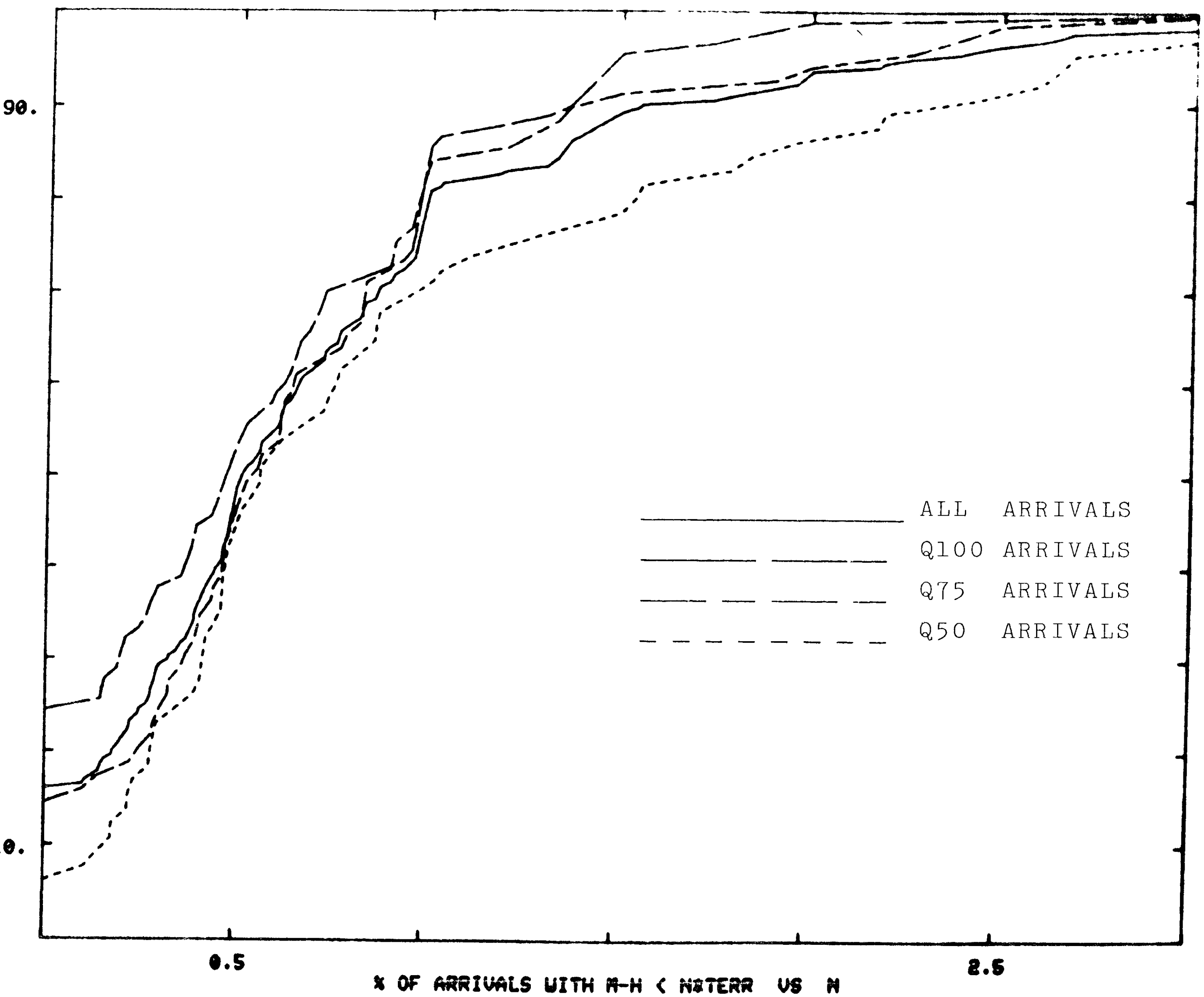


EXHIBIT 6.10

PERCENTAGE OF ARRIVALS WHICH WERE REPICKED

QUALITY	% REPICKED
25	47
50	20
75	8
100	10

EXHIBIT 6.11


90.
10.
0.5
2.5
% OF ARRIVALS WITH M-H < N*TERR VS N
ALL ARRIVALS
Q100 ARRIVALS
Q75 ARRIVALS
Q50 ARRIVALS

7. CONCLUSIONS AND RECOMMENDATIONS

This thesis represents one of the first detailed studies of CEDAR data. Since CEDAR had been in operation for only four months when the data set studied in this thesis was collected the data had a number of problems as discussed in Chapters 2 and 4. The major problem is that the CEDAR data is aliased. Although aliased data is adequate for picking first arrival times, it is inadequate for any detailed analysis of seismic waveforms. One of the most compelling features of CEDAR was that it combined event locations, arrival times and seismograms together in a relatively easy to use package. Having all three types of data together is a unique aspect of the CEDAR database. It is quite unfortunate that the data is aliased. Caltech is currently experimenting with anti-alias filters, and hopefully, the aliasing problem will be resolved soon.

Analysis of the hand picked P wave arrival times shows that beyond about 120 km. arrivals tended to be weak and ambiguous. This increases the possibility of picking late arrival times since the first half cycle or so of the onset may be missed or a later arrival such as a Moho reflection can be picked. Thus if data in this distance range is used for looking for premonitory variations in P delay or VP/VS it should be used with care. Travel time residuals are not Gaussianly distributed as has been suggested by Buland (1976). It is quite likely that there will be one or more larger arrival time errors (greater than one second) for any given event. Thus care must be taken to identify outliers, and a robust location method should be used. The quality assigned to an arrival by an analyst seems to be fairly subjective since several cases were found where weak arrivals were assigned high quality.

The preceding chapter shows that the arrival times determined by the computer system developed in this thesis compare favorably with hand picked arrival times. Of the 400 arrivals picked by both seismologists and the machine, 50 percent agreed to within 0.04 seconds and 75 percent agreed to within 0.12 seconds and the standard error computed for the arrival provides a reasonable measure of the accuracy of the arrival time. Although the picking algorithm is based on one by Stewart (1977) it is much more lenient than his algorithm; Stewart's algorithm is more restrictive in the range of waveforms that it will accept as a seismic arrival. It is designed to accept the larger amplitude implusive arrivals from stations near the event and thus it may miss smaller arrivals or pick them late. The picker used in this thesis is much more lenient in what it will accept as an arrival and the final decision is left up to the location stage. In this way, weaker arrivals can be picked without significantly increasing the false alarm rate.

Thus the real power of the system comes from combining the picking and locating processes. Although a smarter picker would help, it is only the locating stages which has enough information (arrival information from all the sensors) to evaluate the performance of the picking stage. The success of the combined algorithm depends critically on how good the initial picks are, how robust the locator is, how much predictive power an initial location has, and how well the picker can utilize advice (feedback) from the location stage.

In this thesis, the picking algorithm was kept simple, and the robustness of the location algorithm was stressed. The location procedure had to be robust since in a machine environment errors are quite likely. However, this work clearly demonstrates that even with hand picked data a robust location method must be used. This is particularly true if P and S wave arrival times are used in the location. Buland (1976) has shown that using both P and S arrivals stablizes the location process and provides much better depths than using only P arrivals. However, S arrivals are much harder to time accurately than P arrivals since they occur in the P coda and are likely to be contaminated with scattered energy. Thus a robust location method such as BI is definitely required.

The comparison between the robust BI location method and the least squares method showed that the robust method generally provided better location. However, the performance of the robust method can be effected by a poor choice of initial location. The arrival order location method was developed to provide a good initial location. The arrival order location is usually within 3 km and was never more than 14 km. from the final epicenter, where as the first station to report an arrival was typically 11 km. and as much as 90 km. from the epicenter. Another advantage of the arrival order method is that it does not require a velocity model.

Currently, all of the routine analysis of the CEDAR data is done manually. Much of this routine analysis could be done by the computer system developed in this thesis if it were combined with the current CEDAR system. For convenience, we will refer to such a combined system as CEDAR2. Rather than working with the raw CEDAR data, an analyst would review the list of picks made by CEDAR2. Since the initial processing is done automatically, it is easier to produce a more complete and accurate data base of seismic information.

For each event, CEDAR2 would display a portion of each seismogram in the vicinity of the arrival time. For each seismogram, the distance and azimuth of the station as well as the theoretical arrival time and the computer picked arrival time and standard error would be shown as in Exhibits 1.1 and 6.8. The analyst would need to repick only those arrivals which he felt CEDAR2 had done incorrectly. CEDAR2 would automatically recompute the standard error for any new picks.

Since the seismologist is provided with much more information than just the raw seismic data, he is in a better position to make decisions about each arrival. As Pearce and Barley have pointed out, the apparent accuracy of an arrival is often an illusion; having the machine pick and accuracy, as well as the theoretical arrival time, distance and azimuth available, should help an analyst make a better interpretation of the data. Also, it might allow him to recognize that an arrival is more ambiguous than he might have thought without it.

The resulting catalog of seismic parameters will be more accurate and complete. Also, the seismologist is in a good position to make additional measurements which are currently not made routinely. Alternatively, CEDAR2 could be programmed to make them. Such measurements could be quite useful in earthquake prediction. Since a long record of measurements is required in order to distinguish normal variations from anomalous behavior of predictive significance.

Although the results of the picking and locating algorithms are encouraging, they also show that improvements are necessary. One problem is that the method does not always describe the first arrival waveform adequately. Small variations in the first arrival due to noise can alter the computed arrival time and accuracy. Linear filtering can smooth out the effect of noise in some cases, but it could also have the undesirable effect of smoothing a sharp onset. This problem is called a "significance bug" which can manifest itself in either of two forms; either the feature of significance is ignored by processing, or a chunk of noise is incorrectly considered significant.

Techniques used in the fields of Artificial Intelligence and Pattern Recognition may be useful to improve the performance of the algorithm. One promising approach is called syntactic structural analysis of waveforms (Anderson, 1978). This approach has been successful in the analysis of human pulse waveforms (Stockman, 1977; Stockman et al., 1976) and in image processing (Lozano-Perez, 1975). Briefly, structural analysis of a waveform is analogous to the analysis performed by a compiler on a statement of computer programming language. The waveform is broken up into features of morphological significance, or "words". Such features might include peaks, inflection points, segments of constant slope, or relatively flat portions. Grammar rules are then given which describe how these words may be combined to form "sentences". For example, a clipped pulse would be described as "a sharp rising segment preceding a relatively flat segment followed by a sharply falling segment". A glitch could be described as "a sharply rising segment followed by a sharply falling segment". Simple descriptions can be combined to form more complicated descriptions (Stockman, 1977). A seismogram, for example, would be described as "noise followed by seismic signal". Where both "noise" and "seismic signal" would have grammar rules describing their structure.

Current picking algorithms use this type of structural information but in a rather ad hoc manner. Syntactic analysis provides a formalism in which picking algorithms can be easily expressed and compared in common terms (Lozano-Perez, 1975).

Computer-generated descriptions of seismograms should be quite useful for the CEDAR2 system (Anderson, 1976; 1978). They would allow detailed comparison of waveforms to be made which, for more than a few earthquakes, would be tedious to do by hand. For example, Reiter and Monfort (1977) discovered that source - station paths which produced extremely long pulse widths or an unexpected weak arrival preceding the main first arrival crossed a sedimentary trough between the San Andreas and the Vergeles-Zayante faults. This suggests that the waveshape of the first arrival could be used to provide more information about the earth's crust and the effects of local geology on seismograms. The computer could automatically compare waveforms from stations in the same distance range and accumulate a file of anomalous arrivals for further seismological study.

REFERENCES

(Aggarwal et al., 1973) Aggarwal, Y. P., L. R. Sykes, J. Armbruster and M.L.Sbar, Premonitory changes in seismic velocity and prediction of earthquakes, Nature, 241, 101-104, 1973.

(Aki, 1967) Aki, Keiiti, Scaling law of seimic spectrum, J. Geophys. Res., 72, 1217-1231, 1967.

(Aki, 1968) Aki, Keiiti, Seismicity and seismological method, Tectonophysics, 6, 41-58, 1968.

(Aki, 1976) Aki, Keiiti, Signal to noise ratio in seismic measurments, in Volcanoes and tectonosphere ed. by Aoki Hitoshi and Susumu Iizuka (Tokai Univ. Press,Tokyo), 187-192, , 1976.

(Aki and Lee, 1976) Aki, Keiiti, and W. H. K. Lee, Determination of three-dimensional velocity anomalies under a seismic array using first P arrival times from local earthquakes 1. a homogeneous initial model, Bull. Seis. Soc. Amer., 81, 4381-4399, 1976.

(Allen, 1976) Allen, Rex, Automatic earthquake recognition and timing, a paper presented at a Workshop on Automatic and Interactive Processing of Network Data, NCER, Menlo Park, Ca., 1976.

(Ambuter and Solomon, 1974) Ambuter, B. P., and S. C. Solomon, An event-recording system for monitoring small earthquakes, Bull. Seis. Soc. Amer., 64, 1181-1188, 1974.

(Anderson, 1976) Anderson, K. R., Automatic analysis of microearthquake data, a paper presented at a Workshop on Automatic and Interactive Processing of Network Data, NCER, Menlo Park, Ca., 1976.

(Anderson, 1978) Anderson, K. R., Automatic analysis of microearthquake network data, Geoexploration, 16, 159-175, 1978.

(Andrews et al., 1972) Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey, Robust estimates of location: Survey and Advances, Princeton Univ. Press, 1972.

(Andrews, 1974) Andrews, D. F., A robust method for multiple linear regression, Technometrics, 16, 523-531, 1974.

(Anonymous, 1966) Anonymous, Atlas of signals and noise, Teledyne Geotech, Tech. Report No. 66-42, 1966.

(Armstrong and Frome, 1976) Armstrong, R. D., and E. L. Frome, A comparison of two algorithms for absolute deviationcurve fitting, J. Amer. Stat. Assoc., 71, 328-330, 1976.

(Beaton and Tukey, 1974) Beaton, A. E., and J. W. Tukey, The fitting of power series,meaning polynomials, illustrated on band-spectroscopic data, technometrics, 16, 147-185, 1974.

(Ben-Menahem and Bath, 1960) Ben-Menahem, Ari, and Markus Bath, A method for the determination of epicenters of near earthquakes, Geofis. Pura e appl., 46, 37-46, 1960.

(Bolt, 1960) Bolt, B. A., The revision of earthquake epicenters,focal depths and origin-times using a high-speed computer, Geophys. J. Roy. Astron. Soc., 3, 433-440, 1960.

(Bolt, 1970) Bolt, B. A., Earthquake location for small networks using the generalized inverse matrix, Bull. Seis. Soc. Amer., 60, 1823-1828, 1970.

(Bolt, 1976) Bolt, B. A., Abnormal seismology, Bull. Seis. Soc. Amer., 66, 617-624, 1976.

(Borcherdt and Healy, 1968) Borcherdt, R. D, and J. H. Healy, A method for estimating the uncertainty of seismic velocities measured by refraction techniques, Bull. Seis. Soc. Amer., 58, 1769-1790, 1968.

(Brune, 1970) Brune, J. N., Tectonic stress and the spectra of seismic shear waves from earthquakes, J. Geophys. Res., 75, 4997-5009, 1970.

(Buland, 1976) Buland, Ray, The mechanics of locating earthquakes, Bull. Seis. Soc. Amer., 66, 173-187, 1976.

(Capon and Green, 1970) Capon, Jack, and P. E. Green, Statistical accuracy of data used in seismic inversion, _Geophys_. _J_. _Roy_. _Astron_. _Soc_., _21_, 373-386., 1970.

(Cisternas and Jobert, 1977) Cisternas, Armando, and Georges Jobert, Problème inverse des foyers sismiques pour un rèseau local tridimensionnel, _C_. _R_. _Acad_. _Sc_. _Paris_, _284_, 69-72, 1977.

(Claerbout and Muir, 1973) Claerbout, J. F., and Francis Muir, Robust modeling with erratic data, _Geophysics_, _38_, 5, 826-844, 1973.

(Crampin, 1970) Crampin, Stuart, A method for the location of near seismic events using travel-times along ray paths, _Geophys_. _J_. _Roy_. _Astron_. _Soc_., _21_, 535-539, 1970.

(Crosson, 1976a) Crosson, R. S., Crustal structure modeling of earthquake data 1. Simultaneous least squares estimation of hypocenter and velocity parameters, _J_. _Geophys_. _Res_., _81_, 3036-3046, 1976a.

(Crosson, 1976b) Crosson, R. S., Crustal Structure Modeling of earthquake data 2. Velocity structure of the Puget Sound region,Washington, _J_. _Geophys_. _Res_., _81_, 3047-3054, 1976b.

(Dennis and Welsch, 1976) Dennis, J. E., and R. E. Welsch, Techniques for nonlinear least squares and robust regression, _In_ _Amer_. _Statist_. _Assn_.: _Proceedings_ _of_ _the_ _Statist_. _Computing_ _Section_, _83_-_87_, , 1976.

(Dratler and Bakun, 1976) Dratler, Jay Jr., and W. H. Bakun, _Emperical transfer functions for stations in the Central California Seismological Network_, U. S. G. S. Open-file report 76-259, 1976.

(Eaton, 1978) Eaton, J. P., _Frequency response of the USGS short period telemetered seismic system and its suitability for network studies of local earthquakes_, U. S. Geological Survey Open-file Report in preparation., 1978.

(Eaton _et_ _al_., 1970) Eaton, J. P., W. H. K. Lee, and L. C. Pakiser, Use of microearthquakes in the study of the mechanics of earthquake generation along the San Andreas fault in Central California, _Techtonophysics_, _9_, 259-285, 1970.

(Engdahl and Gunst, 1966) Engdahl, E. R., and R. H. Gunst, Use of a high speed computer for the preliminary determination of earthquake hypocenters, Bull. Seis. Soc. Amer., 56, 325-336, 1966.

(Engdahl and Lee, 1976) Engdahl, E. R., and W. H. K. Lee, Relocation of local earthquakes by seismic ray tracing, J. Geophys. Res., 81, 4400-4406, 1976.

(Evernden, 1969) Evernden, J. F., Precision of epicenters obtained by small numbers of world-wide stations, Bull. Seis. Soc. Amer., 59, 1365-1398, 1969.

(Flinn, 1960) Flinn, E. A., Local earthquake location with an electronic computer, Bull. Seis. Soc. Amer., 50, 467-470, 1960.

(Flinn, 1965) Flinn, E. A., Confidence regions and error determinations for seismic event location, Rev. Geophys. Space Phys., 3, 157-185, 1965.

(Freedman, 1966a) Freedman, H. W., The "little variable factor" a statistical discussion of the readings of seismograms, Bull. Seis. Soc. Amer., 56, 593-604, 1966a.

(Freedman, 1966b) Freedman, H. W., A statistical discussion of Pn residuals from explosions, Bull. Seis. Soc. Amer., 56, 677-695, 1966b.

(Freedman, 1968) Freedman, H. W., Seismological measurements and measurement error, Bull. Seis. Soc. Amer., 58, 1261-1271, 1968.

(Freuder, 1976) Freuder, E. C., Affinity: a relative approach to region finding, Computer Graphics and Image Processing, 5, 254-264, 1976.

(Fu, 1974) Fu, K. S., Syntactic methods in pattern recognition, Academic Press, New York, N. Y., 1974.

(Fu, 1977) Fu, K. S. ed., Syntactic pattern recognition applications, Springer-Verlag, New York, N. Y., 1977.

(Garaz et al., 1977) Garaz, Tomås, Cinna Lomnitz, and C. R. de Velosco, An interactive epicenter location procedure for the RESMAC seismic array: I, Bull. Seis. Soc. Amer., 67, 1577-1586, 1977.

(Geiger, 1910) Geiger, L., Herdbestimmung bei erdbeben aus den ankunftszeiten, K. Gesell. Wiss. Gott., 4, 331-349, 1910.

(Gershanik, 1973) Gershanik, Simón, Earthquake location based on the gradient method and minima of beams of directions, Bull. Seis. Soc. Amer., 63, 1829-1840, 1973.

(Greensfelder, 1965) Greensfelder, R. W., The Pg -- Pn method of determining depth of focus with applications to Nevada earthquakes, Bull. Seis. Soc. Amer., 55, 391-403, 1965.

(Gutdeutsch and Aric, 1977) Gutdeutsch, R., and K. Aric, Determination of focal coordinates of near earthquakes in an area of unknown crustal parameters, Pure Appl. Geophys., 115, 667-673, 1977.

(Hadley, 1962) Hadley, G., Linear programming, Addison-Wesley, Reading, Ma., 1962.

(Helstrom, 1960) Helstrom, C. W., Statistical theory of signal detection, Pergamon Press, New York, 1960.

(Hill and Holland, 1974) Hill, R.W, and P. W. Holland, A monte carlo study of two robust alternatives to least squares regression estimation, MIT Center for Computational Research in Economics and Management Science Working Paper No. 58, 1974.

(Hill, 1977) Hill, R. W., Robust regression when there are outliers in the carriers, Harvard Univ. Ph. D. Thesis, 1977.

(Horowitz, 1977) Horowitz, S. L., Peak recognition in waveforms, in Syntactic pattern recognition applications, ed. by K. S. Fu (Springer-Verlag,New York), , 1977.

(Huber, 1972) Huber, P. J., Robust Statistics: a review, Ann. Math. Stat., 43, 4, 1041-1067, 1972.

(Husebye, 1966) Husebye, E. S., A rapid graphical method of epicenter location, Gerlands Beiträge zur Geophysik, 75 Heft 5, 383-392, 1966.

(ISC, 1976) ISC, Bulletin of the international seismological centre, 13, 1, 1976.

(James et al., 1969) James, D. E., I. S. Sacks, E. Lazo L., and P. Aparicio G., On locating local earthquakes using small networks, Bull. Seis. Soc. Amer., 59, 3, 1201-1212, 1969.

(Jeffreys, 1932) Jeffreys, Harold, An alternative to the rejection of observations, Proc. Roy. Soc. Ser. A, 137, 78-87, 1932.

(Jeffreys, 1948) Jeffreys, Harold, Theory of probability, 2nd Ed., Cambridge Univ. Press, Oxford, p. 189-192, 1948.

(Jeffreys, 1962) Jeffreys, Harold, The earth, 4th Ed., Cambridge Univ. Press, Oxford, p. 93-95, 1962.

(Jeffreys, 1973) Jeffreys, Harold, On travel times in Seismology,in, Collected papers of Sir Harold Jeffreys on geophysics and other sciences, Gordon and Breach Science Pub., N. Y., 1973.

(Johnson, 1978) Johnson, Carl, CEDAR - An approach to the computer automation of short period local seismic networks, AGU abs. EOS, 59, 316, 1978.

(Kanamori and Hadley, 1975) Kanamori, Hiroo, and David Hadley, Crustal structure and temporal velocity change in Southern California, Pageoph, 113, 257-280, 1975.

(Kernighan and Plauger, 1976) Kernighan, B. W., and P. J. Plauger, Software tools, Addison-Wesley, Menlo Park, 338 pp., 1976.

(Kerr, 1978) Kerr, R. A., Earthquakes: prediction proving elusive, Science, 200, 419-421, 1978.

(Lee et al., 1972) Lee, W. H. K., R. E. Bennett, and K. L. Meagher, A method of estimating magnitude of local earthquakes from signal duration, U.S.G.S. Open-file report, 75-311, 1972.

(Lee and Lahr, 1975) Lee, W. H. K., and J. C. Lahr, HYPO71 (revised): A computer program for determining hypocenter magnitude, and first motion pattern of local earthquakes, U. S. G. S. Open-file report 75-311, 1975.

(Lindh et al., 1978) Lindh, Allan, Gary Fuis, Constance Mantis, Seismic amplitude measurements suggest foreshocks have different focal mechanisms than aftershocks, Science, 201, 56-59, 1978.

(Lindh et al., 1978) Lindh, A. G., D. A. Lockner, and W. H. K. Lee, Velocity anomalies: an alternative explanation, Bull. Seis. Soc. Amer., 68, 721-734, 1978.

(Lomnitz, 1977a) Lomnitz, Cinna, A fast epicenter location program, Bull. Seis. Soc. Amer., 67, 425-431, 1977a.

(Lomnitz, 1977b) Lomnitz, Cinna, A procedure for eliminating the indeterminacy in focal depth determinations, Bull. Seis. Soc. Amer., 67, 533-535, 1977b.

(Lomnitz, 1978) Lomnitz, Cinna, Reply, Bull. Seis. Soc. Amer., 68, 857, 1978.

(Lozano-Perez, 1975) Lozano-Perez, Tomas, Parsing intensity profiles, MIT Artificial Intelligence Lab. Memo No. 329, Cambridge, Ma., 1975.

(Marr, 1975) Marr, David, The purpose of low-level vision, MIT Artificial Intelligence Lab. Memo No. 324, Cambridge, Ma., 1975.

(Marquardt, 1963) Marquardt, D. W., An algorithm for least-squares estimation of nonlinear parameters, J. Soc. Ind. Appl. Math., 11, 431-441, 1963.

(Marquardt, 1970) Marquardt, D. W., Generalized inverses,Ridge Regression, Biased Linear Estimation, and nonlinear estimation, Technometrics, 12, 591-611, 1970.

(Marquardt, 1974) Marquardt, D. W., Discussion No. 2, Technometrics, 16, 189-192, 1974.

(McCowan and Lacoss, 1978) McCowan, D.W., and R.T. Lacoss, Transfer functions for the seismic research observatory seismograph system, _Bull_. _Seis_. _Soc_. _Amer_., _68_, 2, 501-512, 1978.

(McNally and McEvilly, 1977) McNally, K. C., and T. V. McEvilly, Velocity contrast across the San Andreas fault in central California: small-scale variations from P wave nodal plane distortions, _Bull_. _Seis_. _Soc_. _Amer_., _67_, 1565-1576, 1977.

(Miller and Harding, 1972) Miller, Webb, and S. T. Harding, Error analysis of a five-station P-wave location technique, _Bull_. _Seis_. _Soc_. _Amer_., _62_, 1073-1077, 1972.

(Mityashev, 1965) Mityashev, B. N., _The determination of the time position of pulses in the presence of noise_, Macdonald, London, 190 pp., 1965.

(Miyamura, 1960) Miyamura, Setumi, Local earthquakes in Kii Peninsula, Central Japan. part 4, _Bull_. _Earthquake Res_. _Inst_., _38_, 77-112, 1960.

(Neumann, 1966) Neumann, Frank, _Principles underlying the interpretation of seismograms_, U. S. C. G. S. Spec. Publ. No. 254, 1966.

(Nordquist, 1962) Nordquist, J. M., A special-purpose program for earthquake location with an electronic computer, _Bull_. _Seis_. _Soc_. _Amer_., _52_, 431-437, 1962.

(Northrop, 1970) Northrop, John, Accuracy of earthquake epicenters on the Gorda Ridge, _Bull_. _Seis_. _Soc_. _Amer_., _60_, 265-267, 1970.

(Pakiser and Steinhart, 1964) Pakiser, L. C., and J. S. Steinhart, Explosion seismology in the western hemisphere, _in Research in Geophysics_, _ed_. _H_. _Odishaw_, MIT Press, Cambridge, Ma., 1964.

(Pearce and Barley, 1977) Pearce, R. G., and B. J. Barley, The effect of noise on seismograms, _Geophys_. _J_. _Roy_. _Astron_. _Soc_., _48_, 543-547, 1977.

(Peters and Crosson, 1972) Peters, D. C., and R. S. Crosson, Application of prediction analysis to hypocenter determination using a local array, Bull. Seis. Soc. Amer., 62, 775-788, 1972, 1972.

(Rabiner et al., 1975) Rabiner, L. R., M. R. Sambur, and C. E. Schmidt, Applications of a nonlinear smoothing algorithm to speech processing, IEEE Trans, ASSP-23, 552-557, 1975.

(Reiter and Monfort, 1977) Reiter, Leon, and M. E. Monfort, Variation in initial pulse width as a function of anelastic properties and surface geology in Central California, Bull. Seis. Soc. Amer., 67, 1319-1338, 1977.

(Richter, 1958) Richter, C. F., Elementary seismology, W. H. Freedman and Co., San Francisco, 1958.

(Robinson and Ellsworth, 1974) Robinson, R., R. L. Wesson and W. L. Ellsworth, Variation of P-wave velocity before the Bear Valley California earthquake of 24 February 1972, Science, 184, 1281-1283, 1974.

(Roller and Healy, 1963) Roller, J.C, and J. H. Healy, Seismic-refraction measurements of crustal structure between Santa Monica Bay and Lake Mead, J. Geophys. Res., 68, 5837-5849, 1963.

(Semenov, 1969) Semenov, A. N., Variations in the travel times of transverse and longitudinal waves before violent earthquakes, Bull. Acad. Sci. USSR, Phys. Solid Earth, 3, 245-248, 1969.

(Shannon and Weaver, 1972) Shannon, C. E., and W. Weaver, The mathematical theory of communication, Univ. Ill. Press, Urbana, 1972.

(Simon, 1972) Simon, R. B, Earthquake interpretations, Colorado School of Mines, Golden, Co., 1972.

(Shlien, 1972) Shlien, S., Automatic classification of seismic detections from large aperture seismic arrays, MIT Ph. D. Thesis, Cambridge, Ma., 1972.

(Smith, 1976) Smith, E. G. C., Scaling the equations of condition to improve conditioning, Bull. Seis. Soc. Amer., 66, 2075-2081, 1976.

(Smith, 1978) Smith, E. C., Comments on "A fast epicenter location program", Bull. Seis. Soc. Amer., 68, 845-846, 1978.

(Steppe et al., 1977) Steppe, J. A., W. H. Bakun, and C. G. Bufe, Temporal Stability of P-velocity anisotropy before eathquakes in central California, Bull. Seis. Soc. Amer., 67, 1075-1090., 1977.

(Stevenson, 1976) Stevenson, P. R., Microearthquakes at Flathead Lake,Montana: a study using automatic earthquake processing, Bull. Seis. Soc. Amer., 66, 61-80, 1976.

(Stewart, 1977) Stewart, S. W., Real-time detection and location of local seismic events in Central California, Bull. Seis. Soc. Amer., 67, 433-452, 1977.

(Stockman et al., 1976) Stockman, George, L. N. Kanal, and M. C. Kyle, Structural pattern recognition of cartoid pulse waves using a general waveform parsing system, Comm. A. C. M., 19, 688-695, 1976.

(Stockman, 1977) Stockman, George, A problem-reduction approach to the linguistic analysis of waveforms, Univ. Maryland Computer Sci. Tech. Rep. No. 538, College Park, Md., 1977.

(Tomek, 1975) Tomek, Ivan, More on piecewise linear approximation, Computers and Biomed. Res., 8, 568-572, 1975.

(Torrieri, 1972) Torrieri, D. J., The uncertainty of pulse position due to noise, IEEE Trans. Aero. Elec. Sys., AES-8 NO. 5, 661-668, 1972.

(Tukey, 1977) Tukey, J. W, Explotatory data analysis, Addison-Wesley, Reading Mass, 1977.

(Tversky and Kahneman, 1974) Tversky, A., and D. Kahneman, Judgement under uncertainty,heuristics and biases, Science, 185, 1124-1131, 1974.

(Van Trees, 1971) Van Trees, H. L., Detection,estimation, and modulation theory Part III, John Wiley & Sons, Inc., New York, 626 pp., 1971.

(Welsch, 1975) Welsch, R. E., Confidence regions for robust regression, MIT Center for Computational Research in Economics and Management Science Working Paper No. 111, 1975.

(Welsch and Becker, 1975) Welsch, R. E., and R. A. Becker, Robust nonlinear regression using the dogleg algorithm, MIT Center for Computational Research in Economics and Management Science Working Paper No. 76, 1975.

(Wiggins, 1976) Wiggins, R. A., Interpolation of digitized curves, Bull. Seis. Soc. Amer., 66, 2077-2081, 1976.

(Whitcomb, 1977) Whitcomb, J. A., Earthquake predicition-related research at the seismological laboratory, California Instititute of Technology, 1974-1976, J. Phys. Earth, 25, Suppl., S1-S11, 1977.

(Wood, 1970) Wood, W. A., Transition Network Grammars for Natural Language Analysis, Comm. A. C. M., 13, 591-606, 1970.

(Woodward, 1957) Woodward, P. M., Probability and information theory, with application to radar, Gergamon Press, New York, 1957.