STATISTICAL CLUSTERING OF MICROSEISMIC EVENT SPECTRA TO IDENTIFY SUBSURFACE STRUCTURE

by

Deborah Kay Fagan

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Geophysics

Boise State University

August 2012

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Deborah Kay Fagan

Thesis Title:	Statistical Clustering of Microseismic Event Spectra to Identify Subsurface Structure
Date of Final Oral Examination:	15 June 2012

The following individuals read and discussed the thesis submitted by student Deborah Kay Fagan, and they evaluated her presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Kasper van Wijk, Ph.D.	Chair, Supervisory Committee
John R. Pelton, Ph.D.	Member, Supervisory Committee
Ludmila Adam, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Kasper van Wijk, Ph.D., Chair of the Supervisory Committee. The thesis was approved for the Graduate College by John R. Pelton, Ph.D., Dean of the Graduate College.

ACKNOWLEDGMENT

Many thanks to the following, who without love and support this work would not be possible: my husband, Jason Fuller – you are my best friend and my physical and emotional rock. My in-laws, Shelley and Corey Fuller – for your caregiving from by bike wreck through to watching the wee one so I could work. The Fagan clan – for your emotional support through the trying times. Dale Anderson – for your mentoring, friendship, and meddling throughout my career.

Thank you to the Geophysics faculty at Boise State University, with special thanks to: Dr. Kasper van Wijk, my advisor and teacher – thank you for taking a chance and hanging in there with me for the long haul. Dr. Mila Adam and Dr. Jack Pelton – for not only being excellent teachers, but also being on my thesis committee and providing excellent guidance and suggestions.

Thank you to my student colleagues at Boise State University: Mike Thoma and Brady Johnson – for putting up with me in the office and watching Mags on occasion. I think I still owe one of you a bottle of fine Washington wine. Mike Thoma, Shaun Finn, Thomas Blum, Dylan Mikesell, Andy Lamb, Shannon Murray, and Emily Hinz – for your study suggestions and help on the homework. Additional thanks to Andy Lamb for providing one of the figures in this thesis and the associated journal manuscript and Shannon Murray for being my class project partner and swim

buddy.

Thank you to Ian Bradford at Schlumberger CRS for supporting a portion of my educational costs and providing guidance on this research. Thank you to Jim Rutledge at MEQ, Inc. for providing data, knowledge, and guidance on this research and co-authoring the associated journal manuscript.

AUTOBIOGRAPHICAL SKETCH

I'm a born and raised Montana girl, graduating from Billings Senior High School in 1987 and Montana State University in Bozeman in 1992 and 1995 with a BS in Math and an MS in Statistics, respectively. I joined several other Montana natives at Neptune & Company in Los Alamos, NM out of school and worked for over five years on environmental remediation projects for US DoD and DOE. Longing for my family, better flyfishing, and a broader range of work, I relocated to Richland, WA and Pacific Northwest National Laboratory in 2001. In seven years there, my career exposure expanded to include nuclear physics and seismology. I liked seismology so much I decided to return to graduate school for a PhD in geophysics at Boise State University. After deciding to go back to school, but prior to actually doing so, I met the guy who is now my husband. He stayed in Richland with my dog while I ventured off to Boise to follow my dream. Dreams change, however, when living far away from one's husband and facing a loudly ticking biological clock. We became the proud parents of daughter Deena, born two weeks after finals my second year of school. After taking a year off to spend with Deena and consulting with my advisor at BSU, I decided to finish a Master's degree in geophysics, long-distance. This thesis is the culmination of that work.

ABSTRACT

Understanding subsurface structure by studying microseismicity influences a wide range of activities, including energy extraction, aquifer storage, carbon sequestration, and seismic hazard assessment. Identifying individual fractures in a larger fault system is key to characterizing, understanding, and potentially mitigating risks of natural or induced seismicity.

A year-long study associated with a carbon dioxide (CO_2) sequestration project was conducted at the Aneth oil field in southeast Utah to record microseismicity at a single downhole geophone array. A previous analysis located events by first identifying event multiplets consisting of highly correlated time-domain waveforms on receivers shallower than the depth of the microseismic events. Then, a relative location algorithm was used within each multiplet. Hypocenters turned out to be in a layer not directly impacted by either water or CO_2 injection or oil extraction. Nevertheless, the locations outlining faults are consistent with the geology of the basin.

In this thesis, hierarchical agglomerative clustering is used to identify subtle differences for one multiplet at the deepest receiver in the array, whose waveforms might include guided waves. Each event starts out as its own cluster, after which events are iteratively combined based on a dissimilarity metric until a single, final cluster results. Two distance measures are defined, spectral and temporal distance, and used to calculate dissimilarity in the clustering algorithm.

While time-domain clustering was inconclusive, clustering in the frequency domain reveals first spectral differences between two groups of events in multiplet 18, which may originate in different lithologies. A more detailed look identifies subclusters in one of these groups that organize spatially. Subtle spectral differences are detected that are not the result of attenuation and may identify individual *en echelon* fractures within the same lithological unit.

More investigation into the application of hierarchical agglomerative clustering to event spectra and waveforms is needed to identify geophysical conditions where the method could be further utilized. Additional station components, stations, and multiplet analysis could further characterize the method's strengths and constraints, as well as refinement of the geophysical interpretation of results.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1:

INTRODUCTION

Monitoring induced seismicity provides valuable information regarding fluid transport properties in hydrocarbon (Maxwell and Urbancic, 2001) and geothermal (Saar and Manga, 2003) reservoirs, geomechanical effects in carbon sequestration (Streit *et al.*, 2005), and hydraulic properties in aquifer recharge and recovery. Besides classic location algorithms based on the arrival of different wave phases, more advanced techniques have been developed based on differences between closely spaced events (Waldhauser and Ellsworth, 2000; Zhang and Thurber, 2003). Beyond hypocenter estimation, classical time series methods have been used to estimate geomechanical model parameters (Christiansen *et al.*, 2005; Saar and Manga, 2003), while coda wave and spectral methods can extract information about velocity changes in the medium (Haney *et al.*, 2009; Payan *et al.*, 2009; Snieder and Hagerty, 2004; Zadler *et al.*, 2005). Furthermore, waveform correlation methods have been utilized to detect low-magnitude and microseismic events (Schaff and Beroza, 2004; Schaff and Richards, 2004, 2011), refine phase arrivals and event hypocenters (Hansen *et al.*, 2006; Rowe *et al.*, 2002; Snieder and Vrijlandt, 2005; Song *et al.*, 2010), determine event focal mechanisms (Hansen *et al.*, 2006), and to classify unknown source events (Harris, 1991; Shumway, 2003; Shumway and Stoffer, 2006).

In energy extraction operations, identifying individual faults, fault systems, and associated structures influence the location and depth of production wells to ensure the correct fluid conducting faults are tapped. Similarly, location of microseismicity is important for injection wells to ensure that locations, depths, and injection rates will generate fractures in desired fluid-bearing layers. Injection well locations and injection rates in carbon sequestration operations are controlled by fault system structure in managing gas migration and ensuring cap rock integrity. Seismic hazard assessment is affected by fault characterization, since longer faults have the potential to increase seismic hazard over smaller ones, thus influencing surface building and development.

Presented here is an application of statistical methods to analyze subtle differences within clustered events. Events were first analyzed in the time domain to determine whether medium changes over time could be detected. Upon finding no evidence of temporal medium changes, a frequency-domain analysis was performed. Statistical clustering of the power spectra identified evidence of subsurface structure not evident in the time domain. Statistical clustering and correlation analysis of event spectra illustrate how one might infer structure from limited amounts of data and show how location information may be gleaned from a receiver previously omitted from event hypocenter estimation. Chapter 2 discusses the site geology and details of passive microseismic data collection and event location. Details of the analytical methods are discussed in Chapter 3, frequency domain results are presented in Chapter 4, time domain comparisons in Chapter 5, and Chapter 6 presents conclusions and recommendations for future work.

CHAPTER 2:

FIELD BACKGROUND AND MICROSEISMIC DATA COLLECTION

2.1 Geologic Background and Reservoir History

The Aneth unit is one of four oil producing units located within the Greater Aneth oil field of the Paradox Basin in southeast Utah (see Figure 2.1). The lower portion of a stratigraphic column for the Greater Aneth Field from Hintze and Kowallis (2009) is shown in Figure 2.2. Of interest here are the Pennsylvanian and Mississippian subperiods, from which oil and gas are extracted, into which CO_2 is injected, and where microseismicity occurs. The formations of interest, from oldest to youngest, are the Mississippian Leadville (two strata: lower dolostone and upper lime mudstone to peloidal to crinoidal lime wackestone), and Pennsylvanian Molas (poorly stratified silt and sandstones), Pinkerton Trail (alternating thin beds of mudstone and shales interbedded with limestone), and Paradox (cyclic intervals of dolostone, shale, and salt). Within the Aneth Unit, the Paradox Formation has been informally divided into "production zones"; from oldest to youngest, named: Alkali Gulch, Barker Creek, Akah, Desert Creek, and Ismay. Oil is produced from the Desert Creek and Ismay zones, the lowest three are referred to as "Paradox salts" (Carney, 2010).

Structural orientation within the Aneth unit is consistent with that of the Paradox Basin and Greater Aneth (Chidsey, Jr., 2009; Carney, 2010; Morgan *et al.*, 2010). Figure 2.1 shows northwest-southeast oriented fold and fault belt along the northeast portion of the Paradox Basin. Surface deformation bands within Greater Aneth indicate a northwest-southeast trend (see Figure 2.3). Similarly, structure contour maps of Paradox Formation strata identify the Aneth area as a northwest-southeast contour orientation, as seen in Figure 2.4 from Carney (2010), which shows structural contours of a shale layer within the Desert Creek zone. Three-dimensional seismic studies of Greater Aneth show basement faults of Mississippian and Pennsylvanian age that strike northwest-southeast (Rutledge and Soma, 2010). Chidsey, Jr. (2009) reports that the Aneth unit is virtually free of folds and faults, with the exception of a NW-SE striking fault observed in the Ismay and Desert Creek zones of the Paradox formation (see Figure 2.4). Cores reveal evidence of faulting in these zones, but not in the overlying DeChelly or Navajo Sandstones, leading Chidsey, Jr. (2009) to the conclusion that the fault is short with minimal displacement.

This field was discovered in 1956, with water-enhanced oil recovery commencing in 1961. In 2008, the field operator and the Southwest Regional Partnership for CO_2 Sequestration began a carbon sequestration field study at Aneth by converting injection wells from water to gas. As part of the site monitoring plan for the carbon sequestration field study, a permanent, 60-level, down-hole geophone array was installed in a former injection well on the western side of the field (see Figure 2.5). With the conversion to gas-injection, a deep brine injection well was installed to dispose of excess water. The well was drilled approximately 1 km northwest of the geophone array, into Leadville formation limestone to a depth of approximately 2260 m (US EPA Region IX UIC Program, 2007). The well has four lateral arms as shown in Figure 2.5; injection rates varied between 12,000 and 20,000 m^3 of water per week between March, 2008 and March, 2009 (Chidsey, Jr., 2009).

2.2 Microseismic Data Collection and Event Locations

Geophones in the permanent array were set between 805 and 1705 m at 15 m spacing. With a sampling rate of 1500 Hz, the deepest 18 geophones (between 1705 m and 1445 m) are 3-component, the other sensors record the vertical component only. The 3-component geophones have a right-hand three-dimensional coordinate system so that a Z signal is positive for a compression arrival from above.

A sonic log was obtained from the deep injection well before injection began, from which a 1-D velocity model was estimated. Figure 2.7 shows the sonic log, velocity model based on a 100-ft median smooth, and 24 geophone depths. These 24 geophones, which include the deepest 18 3-component receivers and 6 shallower receivers spaced at 107 m, were used for microseismic location analysis.

Microseismic monitoring began at approximately the same time as salt water injection. More than 3800 events were recorded between March, 2008 and March, 2009. For 1212 of these events, P- and S-wave arrival times were obtained from the 24 geophones shown in Figure 2.7. Except for the 8 deepest geophones, which may have recorded waves critically refracted at the top of the Leadville, Rutledge

Figure 2.1: Paradox Basin oil and gas fields (modified from Figure 1-1 in Chidsey, Jr., 2010).The study area is in the Greater Aneth oil field and marked by a red oval.

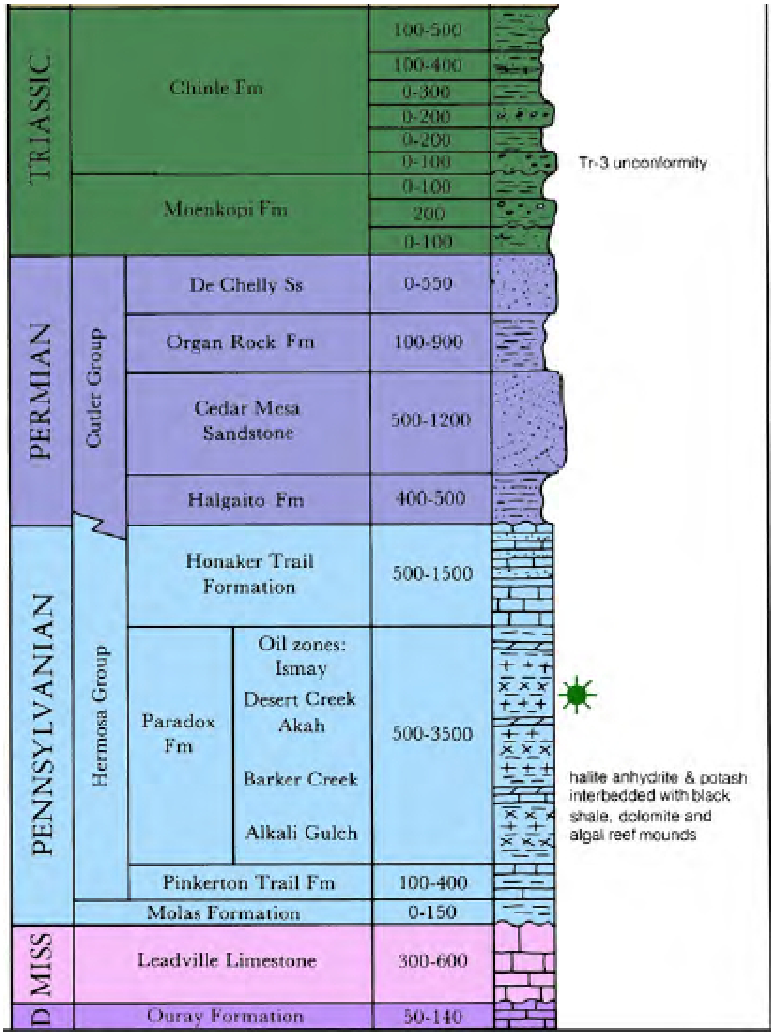


Figure 2.2: The lower portion of a stratigraphic column for Aneth (modified from Hintze and Kowallis, 2009), with formation thicknesses in feet. Oil is produced from the Ismay and Desert Creek zones within the Paradox formation. The lower zones of the Paradox formation (Akah, Barker Creek, and Alkali Gulch) are referred to as Paradox salts, as seen in Figure 2.7.

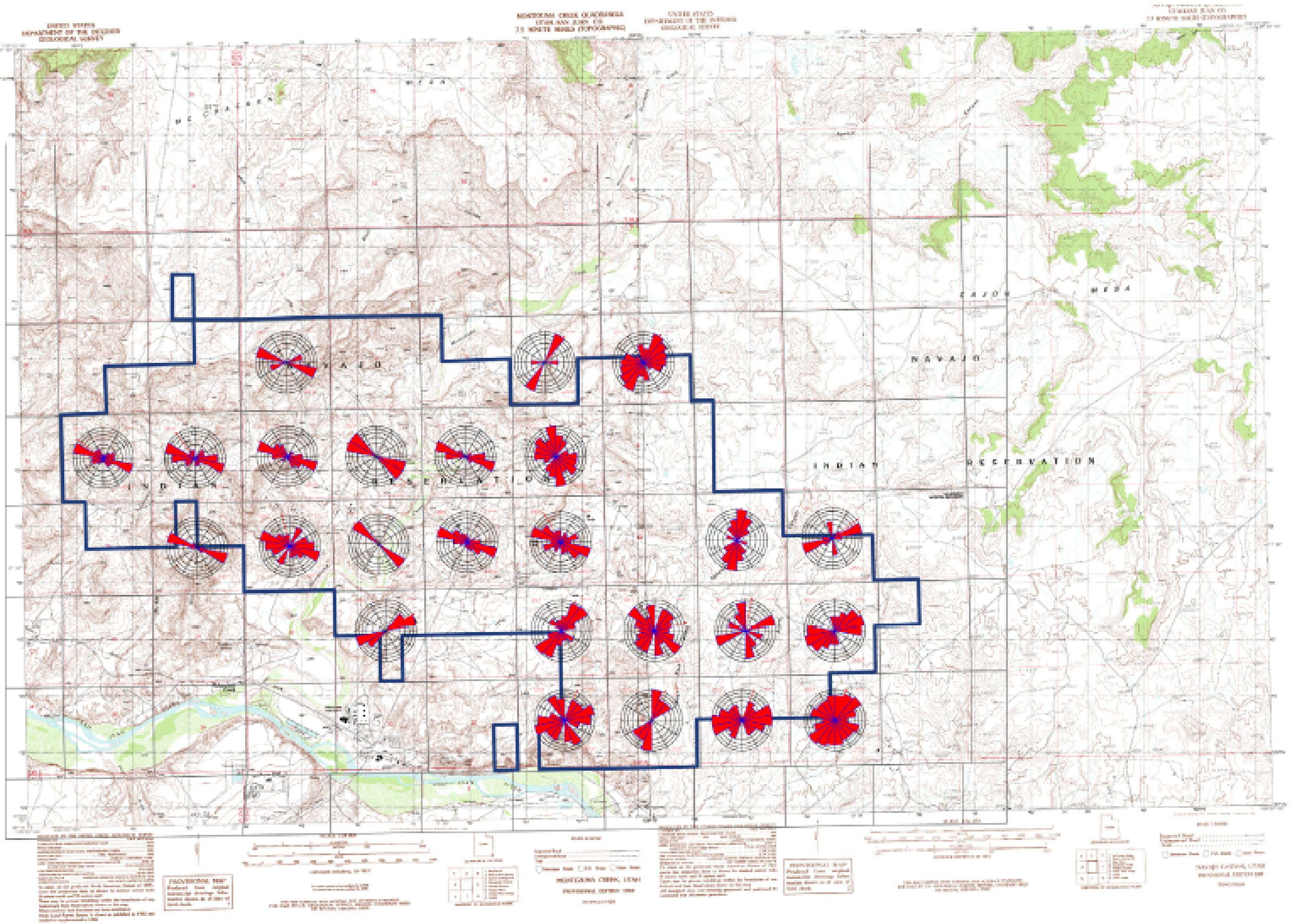


Figure 2.3: A rose diagram of surface deformation bands observed at Aneth (from Figure 2-5 in Morgan *et al.*, 2010). Each rose corresponds to deformation bands observed in each 1 square mile section. Of interest is the southwest quadrant of the field, where orientation is generally NW-SE.

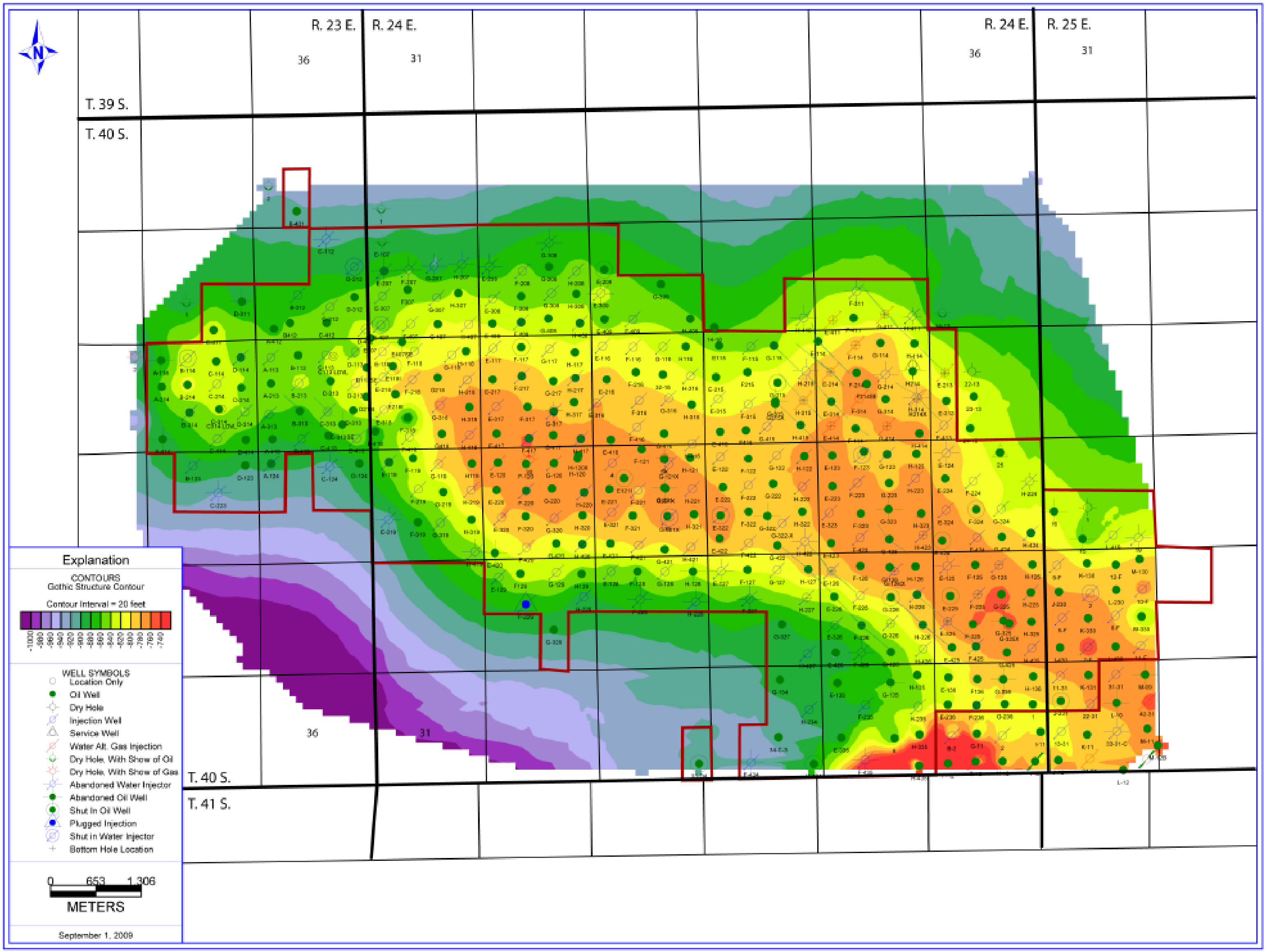


Figure 2.4: A structural contour map of a shale layer within the Desert Creek zone (from Figure 3-8 in Carney, 2010). The Aneth unit is outlined in red. Symbols of interest are solid green dots, marking production wells, and hollow circles with an intersecting diagonal line which identify injection wells.

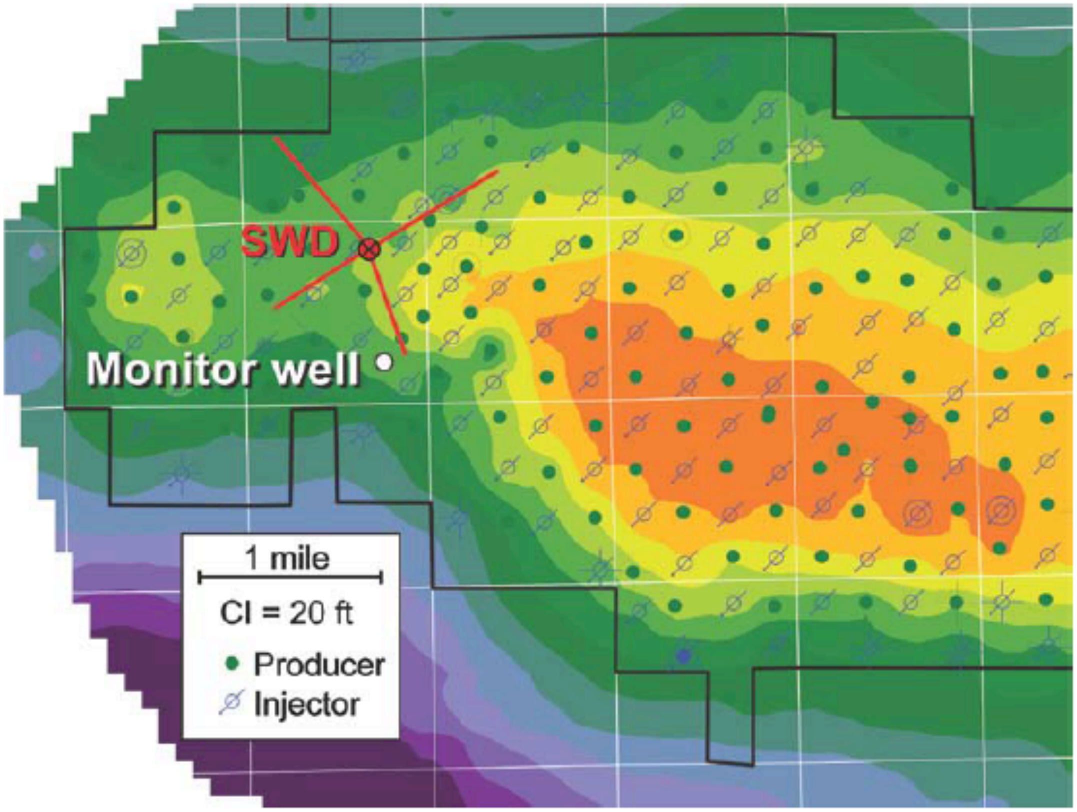


Figure 2.5: Locations of the deep injection well (“SWD”) and geophone array (“Monitor well”) (from Figure 12-1 in Rutledge and Soma, 2010). Approximate depths of the laterals on the injection well are: southeast, 2240 m, northeast, 2198 m, northwest, 2250 m, and southwest, 2206 m (US EPA Region IX UIC Program, 2007).

and Soma (2010) report that P- and S-wave phases tend to be impulsive and easy to identify. Events were initially screened by manually picking P-wave arrivals on waveforms from just a few receivers. These waveforms were then clustered based on waveform similarity (in the time domain) and re-picked using the methods of Rowe *et al.* (2002), which resulted in sets of events with similar wavefields called *multiplets*, using a threshold correlation coefficient of 0.8.

Rowe *et al.* (2002) identify an algorithm used to refine phase picks based on waveform correlation and clustering. Data are first organized based on station-by-station analysis to take advantage of waveforms that have similar source receiver paths. Clustering is then performed over events using waveform correlations across variable window-widths and multiplet composite waveforms are formed via within-cluster stacking. Waveforms are polarization filtered (rotated), energy normalized, and coherency filtered before identifying integer and sub-sample lags of maximum cross correlation, which are then used to estimate phase pick corrections based on an L1 norm solution to a system of equations. A master event is identified for each multiplet, using high signal-to-noise events or composite waveforms and relative, within-cluster epicenters are estimated using a relative location algorithm of choice. The waveform correlation algorithm is then reapplied to the master event waveforms to adjust absolute locations.

Of the 47 multiplets located on the southwest boundary of the field, multiplet 18 with $n = 73$ events, was chosen to research, because of its highly correlated waveforms that cover the 12-month study period and large number of events with high-confidence picks at the geophones used for multiplet identification. The number of P picks for multiplet 18 varies by receiver along the array (as well as by component) from the

3-component receivers. For instance, 70 picks are available from the east component data for the geophone at 1552 m (the 11^{th} geophone from the bottom), whereas 67 picks are available from the deepest receiver (geophone 1).

Rotated east-component data from Geophone 11 for 25 of the 70 events from mulitplet 18 are shown in Figure 2.8. Peak-normalized waveforms starting 0.1 s prior to the P-wave arrival are aligned according to the maximum correlation coefficient over the first 0.25 s. To verify the 0.8 correlation coefficient threshold, starting with the P-wave arrival for each event, the first 0.25 s (generally the region between the P and S arrivals) were cross correlated, with all coefficients exceeding 0.85 but for one. Although this waveform was not well correlated at any receiver along the array, it was not deleted from the analysis since it was included in multiplet 18 by LANL.

Rutledge and Soma (2010) applied a master-slave location scheme to tie the weaker events to the first arrivals of the stronger events. High Signal-to-Noise Ratio (SNR) master events were formed by stacking events in a multiplets or were chosen from a single large magnitude event from a multiplet. As a result, the first arrivals could be reliably picked on this master event. The deepest eight geophones (geophones 1 - 8) are omitted from the location analysis because of P-wave paths critically refracted at the top of the Leadville formation. The left panel of Figure 2.9 from Rutledge and Soma (2010) shows correlated P-wave picks as well as first arrivals for the radial component of the 3-component geophones. The early arrivals at the lower geophones correspond to energy from the critically refracted waves, as seen from the P-wave raypaths in the right panel of this Figure.

For each master event the source locations (radial and depth positions with respect to the vertical receiver array) were determined using an iterative least-squares method to best fit the P- and S-wave arrival times (e.g., Rutledge and Phillips, 2003) plus reflected SV phases (similar to Phillips *et al.*, 1989). The travel time residuals for the master event locations were then applied as station corrections to the remaining events, while omitting the use of any arrivals from the lower eight receivers of the array.

Azimuths to the sources were determined from the horizontal-component P-wave particle motion trajectories (Rutledge and Soma, 2010). The particle-motion data from the array were highly linear and consistent between receiver levels. Azimuths were obtained by averaging over multiple receiver levels. Relative azimuth error was about 1°, measured as the standard error of the mean azimuth. Relative depth errors average 4 m, radial errors average 8 m.

Figure 2.6 shows the 1212 microseismic events located by LANL (Rutledge and Soma, 2010). Events organize into two clusters on opposite sides of the reservoir; the northeast side includes 44 events, the southwest side, 1168 events, which is the focus of discussion and analysis here. Events in the southwest cluster trend along a northwest-southeast fracture zone (consistent with the geologic structure of the area) at least 1500 m long within a depth range of 1830 to 2037 m. Figure 2.10 shows the southwest cluster of events along with the location of the geophone array. Multiplet 18 event depths range between 2002 and 2030 m; the centroid of this multiplet is located approximately 1690 m southwest of the array, 320 m below the deepest receiver and 219 m above the closest lateral of the deep injection well. Assuming no change of formation depths between the salt water disposal well and event locations, multiplet 18 events occur in either the Paradox salt layer or the Pinkerton Trail formation (Figure 2.2). The build-up contours in Figures 2.5 and 2.6 here, as well as additional

Figures in Rutledge and Soma (2010), indicate that the formations are thicker in the region of multiplet 18 than at the disposal well where the sonic log was obtained. Structural contour maps are only available for formations from the Desert Creek and above, but if the structure is consistent in the lower formations where microseismicity is occurring, multiplet 18 events may actually be in the Pinkerton Trail or Molas formations, rather than in the Paradox salt layer. This scenario is more likely, as both the Molas silt and sandstones and Pinkerton Trail mudstones, shales, and limestones will fail before the ductile salt layers in the Paradox Formation.

A Mw 3.7 earthquake (the Bluff earthquake) struck the region on June 6, 2008. It was located approximately 13 km northwest of the geophone array at a depth of 9.5 km. This earthquake was used to verify the azimuths to the events shown in Figure 2.10. The moveout of first arrivals for the earthquake clearly indicate an upward propagating arrival across the array, and given the polarity of the vertical components on the array, indicate dilatational first motion. Similarly, for the microseismic events, which all occur below the bottom of the array, the vertical component at a subset of geophones show dilatational first motion. Horizontal component first motions are all opposite of the Bluff earthquake, putting the microseismic events in the southeast quadrant with respect to the array.

Similar to Figure 2.8, rotated east-component waveforms from the deepest geophone for the same 25 events from multiplet 18 are shown in Figure 2.11 and cross correlation percentiles are shown in Figure 2.12. While these waves are not as similar with one another as those from Geophone 11, they are still highly correlated in the region of the P-wave arrival, with maximum cross correlations ranging between 0.76 and 0.99.

The Nature of the Seismicity at Aneth Field

Rutledge and Soma (2010) study both the space-time development of microearthquake occurrence and the correlation between seismicity and injection/production activities. Using the Gutenburg-Richter law $log(N) = a - bM$, where N is the number of events having magnitude that exceeds M, Rutledge and Soma (2010) fit a value $b = 2.0$. Tectonic seismicity typically results in $b \approx 1$. The high b-value at Aneth indicates a higher proportion of small events to large ones, characteristic of fluid-injection induced seismicity or natural swarm seismicity. Natural swarm seismicity typically precedes volcanic eruption or is associated with the migration of crustal fluids through fracture networks (Lay and Wallace, 1995). A population of earthquakes dominated by small events is generally considered an indication of many small discontinuous faults accommodating strain accumulation. Additionally, larger magnitude events occur near time gaps in the seismicity, suggesting that the subsurface structure is made up of discrete segments where stress may accumulate.

Based on the distance between the salt water injection well and microseismicity (approximately 1200 m), the absence of discernable correlation between injection rates and seismicity, and the long history of injection at Aneth, Rutledge and Soma (2010) conclude that events occurring within the study period were not a direct result of deep salt water injection into the Leadville or CO_2 injection into the Desert Creek zone, but rather stress release related to overall reservoir reduction over the production life of the area.

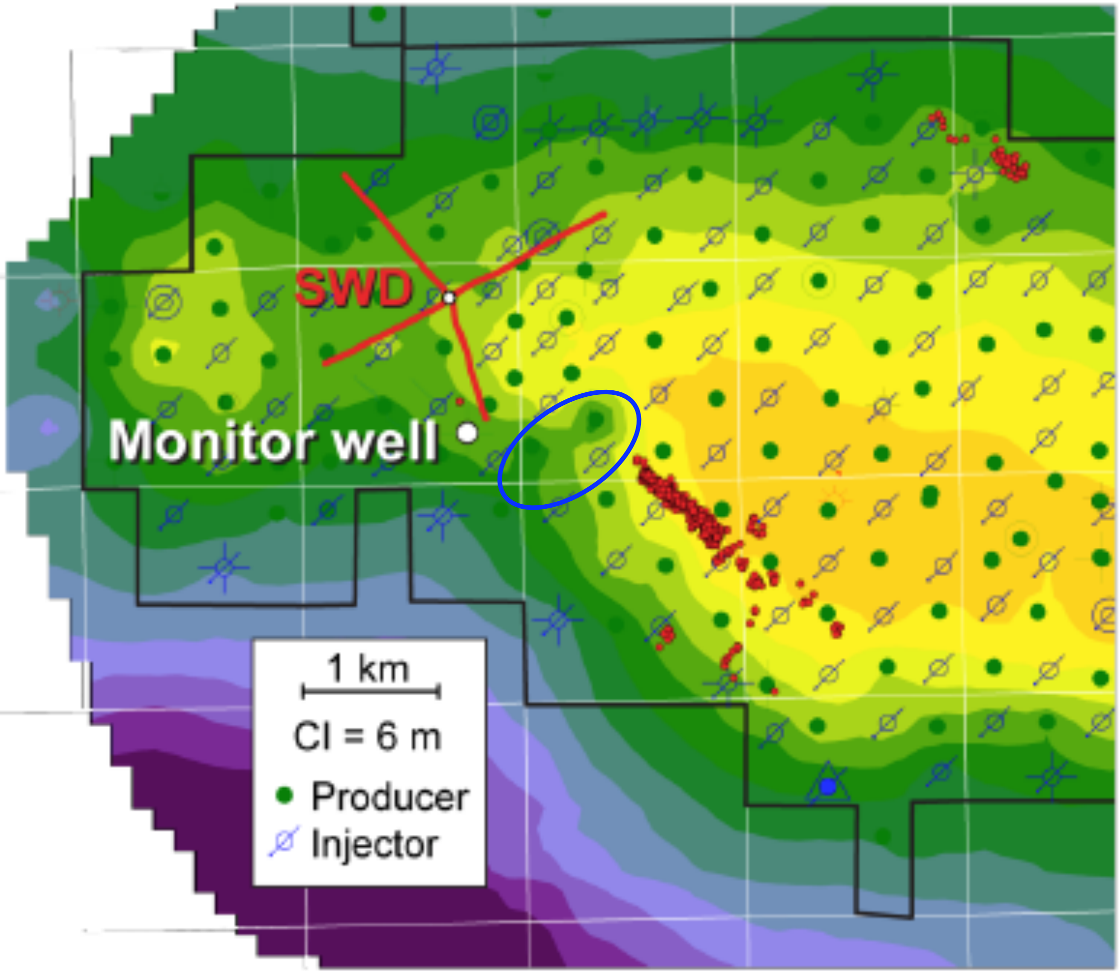


Figure 2.6: Locations for 1212 microseismic events recorded at the geophone array ("Monitor well") in the year March 2008 - 2009 (modified from Figure 12-1 in Rutledge and Soma, 2010). The blue oval shows the approximate location of a small NE-SW striking fault in the Ismay and Desert Creek zones of the Paradox formation believed to have minimal displacement. The thickness of the Desert Creek in the southwest where microseismicity is occurring is 12 - 20 m.

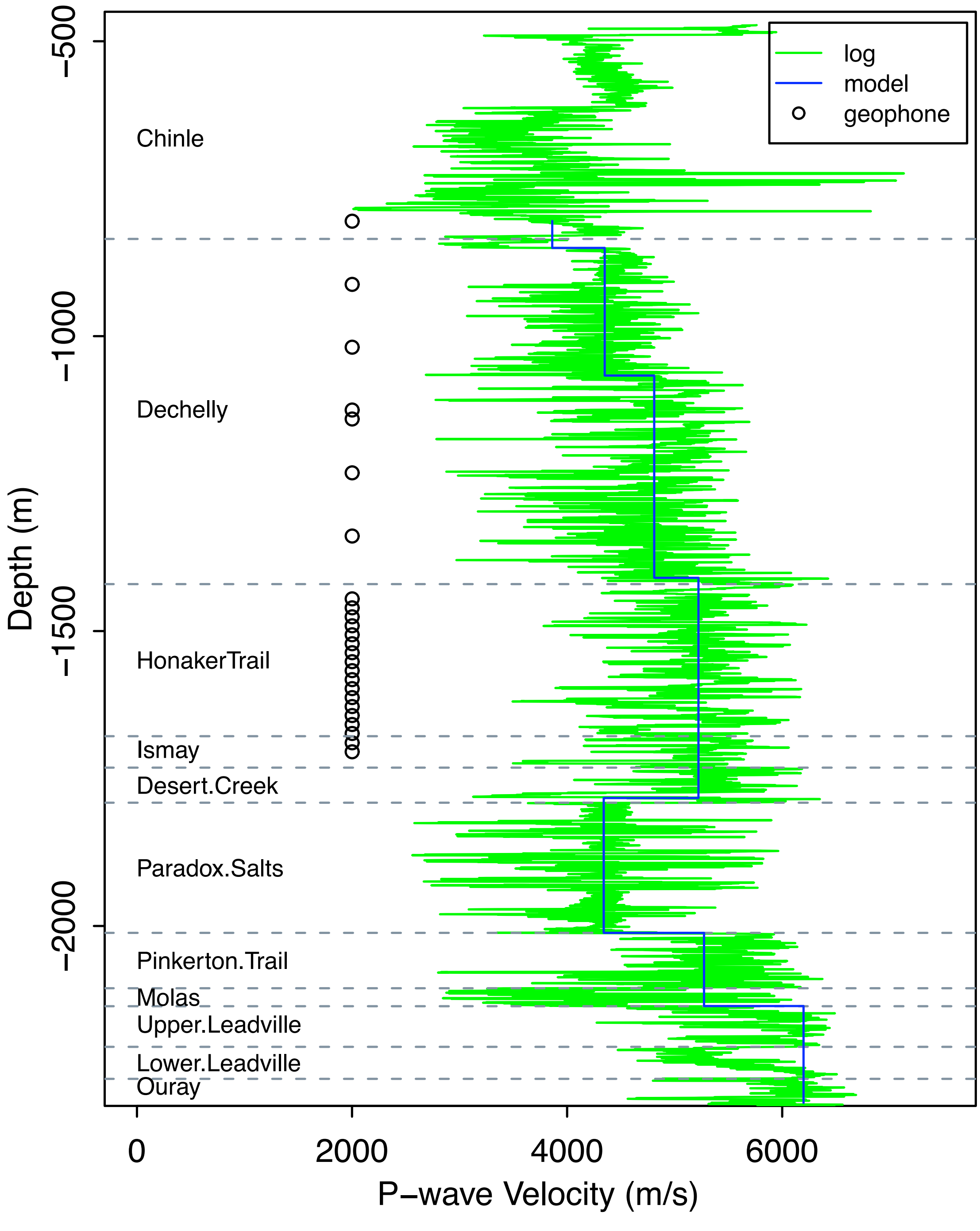


Figure 2.7: Sonic log, P-wave velocity model, and geophone depths used to estimate hypocenters. S-wave velocity is $V_p/1.625$ based on empirical analysis (Rutledge and Soma, 2010). Paradox salts correspond to the Akah, Barker Creek, and Alkali Gulch zones of the Paradox formation in Figure 2.2.

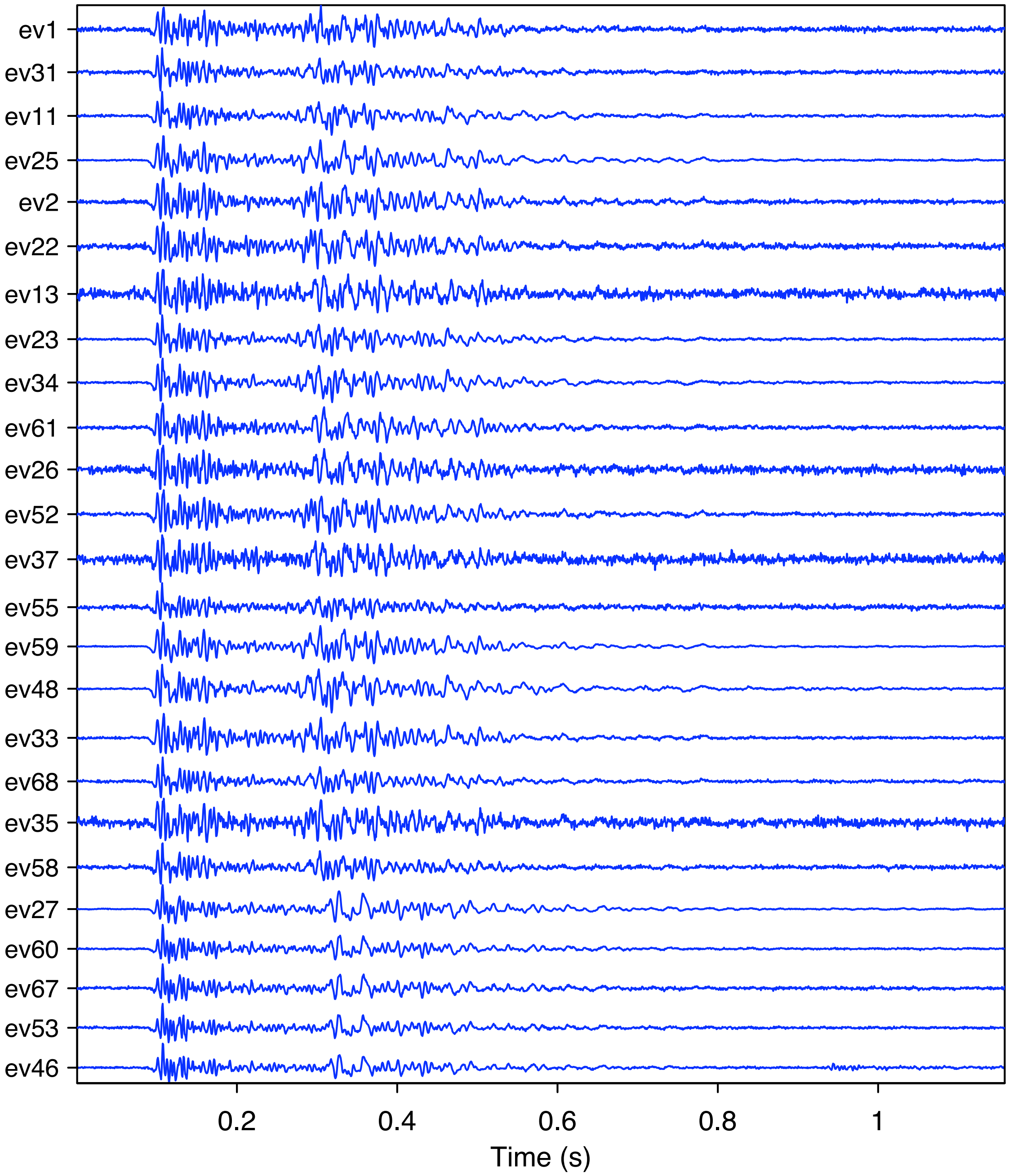


Figure 2.8: Twenty-five waveforms from multiplet 18 from the Geophone 11 (≈ 1552 m). Each waveform starts 0.1 s prior to the P-wave arrival and is normalized by its maximum amplitude. Waveforms are highly similar, as expected by multiplet identification.

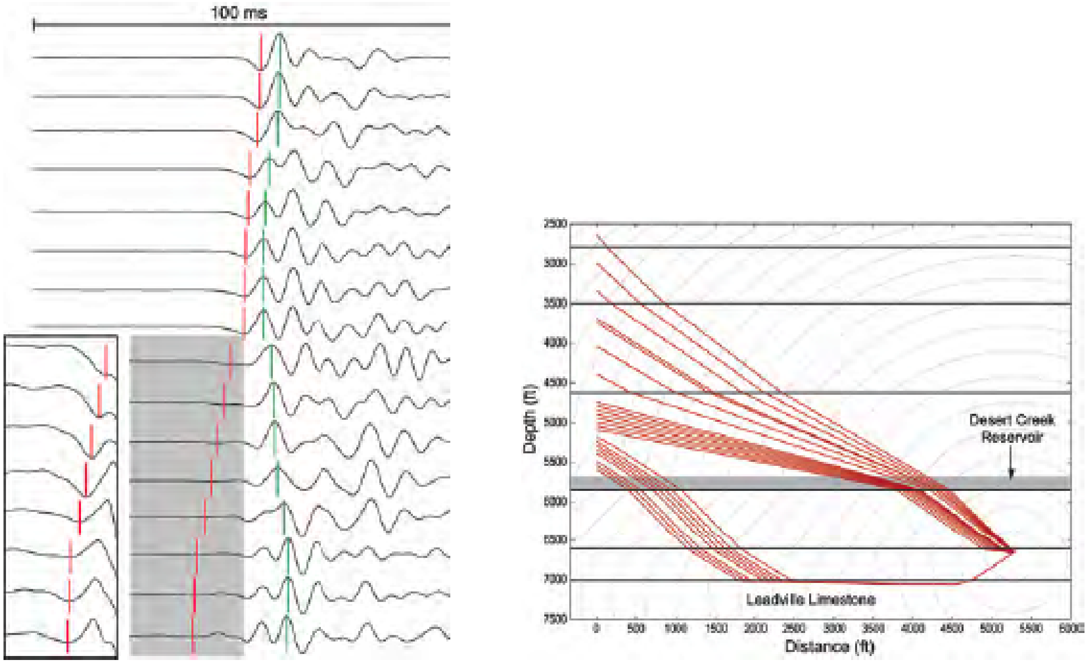


Figure 2.9: P-wave arrivals over the lower portion of the geophone array (left panel) and P-wave raypaths from a master event to the array (right panel) (from Rutledge and Soma (2010)). The green lines in the left panel correspond to the picks based on waveform correlation, while the red lines represent true first arrivals. The lower geophones in the array show weak energy arrivals corresponding to critically refracted waves at the top of the Leadville formation, as shown by the P-wave paths in the right panel.

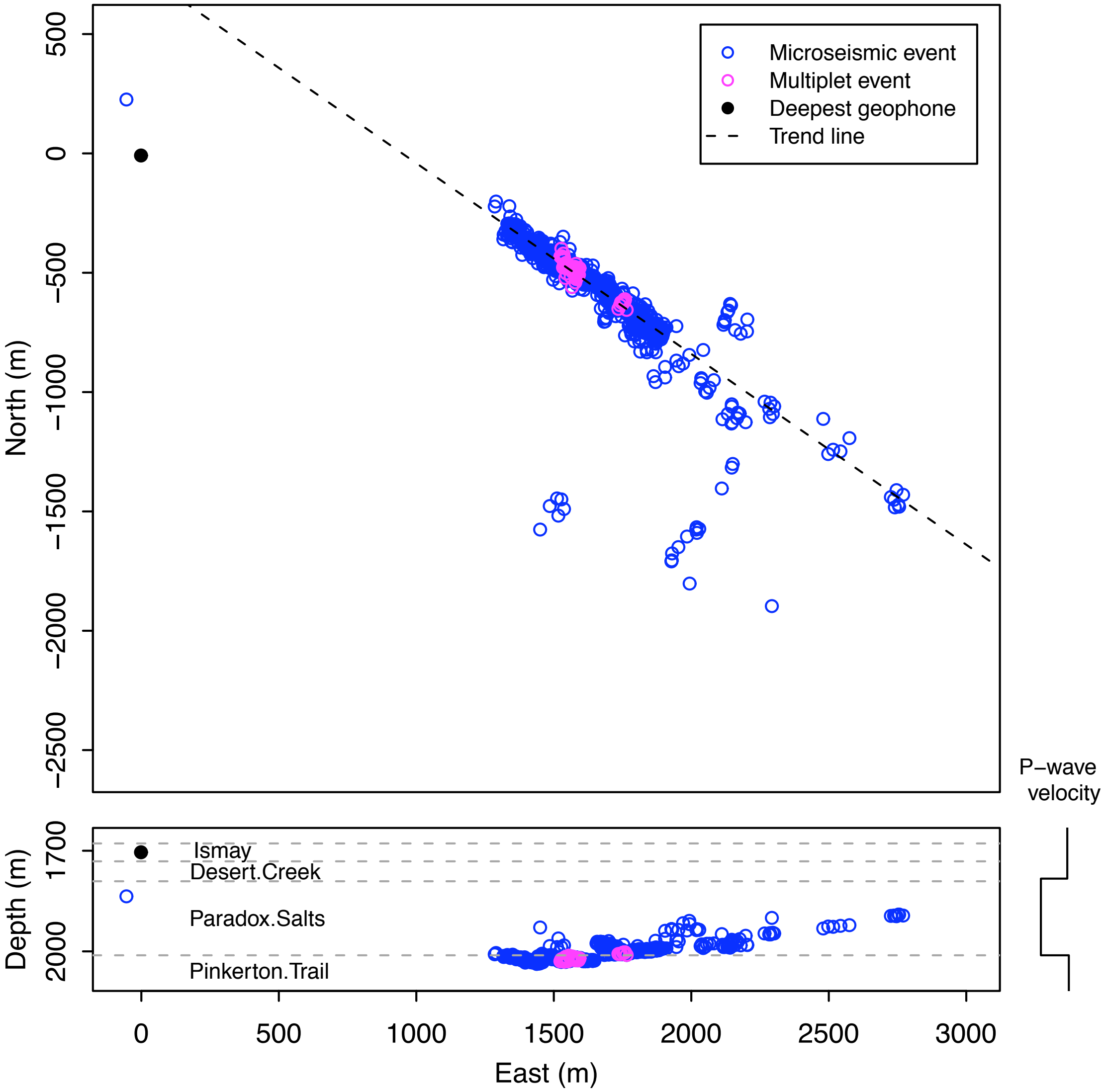


Figure 2.10: The 1168 events located on the southwest boundary of the reservoir, with the geophone array plotted at the origin. Events follow a NW-SE lineament, with slight dip to the west. The 67 magenta events are from multiplet 18, which show similar depths but a geographic separation. P-wave velocity changes are shown to the right of the bottom panel.

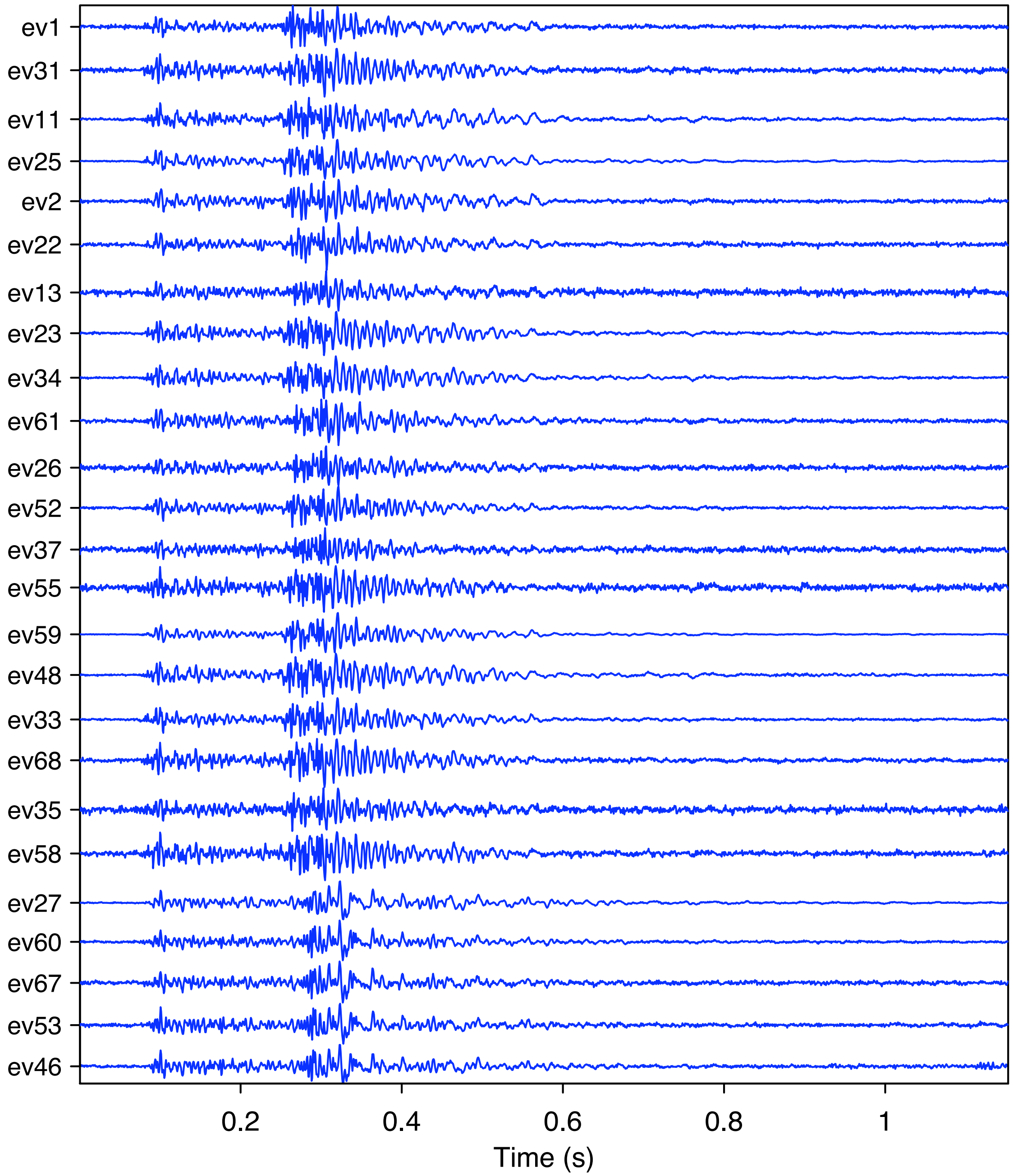


Figure 2.11: Twenty-five waveforms from multiplet 18 from the deepest receiver, aligned by the lag of the maximum correlation coefficient around the maximum amplitude. Each waveform starts 0.1 s prior to the P-wave arrival and is normalized by its maximum amplitude. As noted by Rutledge and Soma (2010), P-wave arrivals for these waveforms are weaker than those from Geophone 11 in Figure 2.8. The bottom five traces can be visually differentiated from the top twenty.

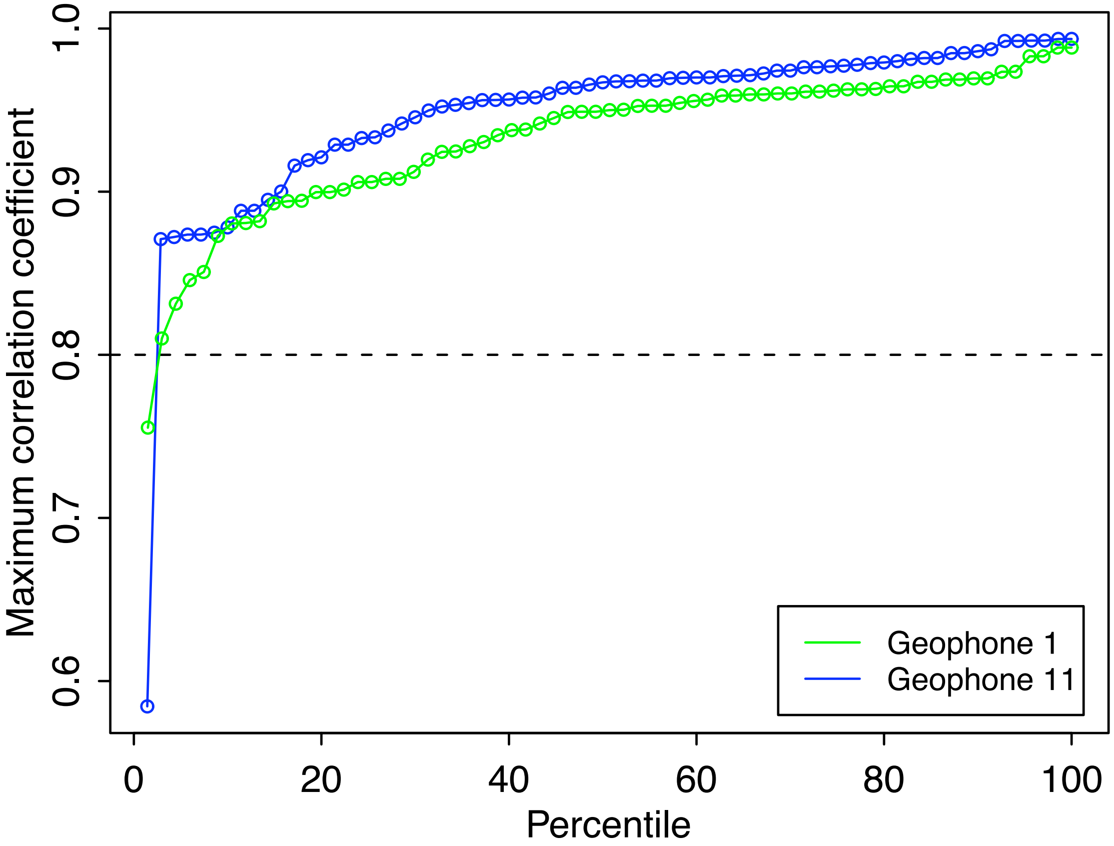


Figure 2.12: Percentiles of the multiplet 18 maximum cross correlations for Geophones 1 and 11 east-component waveforms. Cross correlations were computed for the first 0.25 s of the waveforms, starting with the P-wave arrival. The percentile is the proportion of the cross correlations less than or equal to the ordinate. Geophone 1 correlations are slightly lower than those at Geophone 11, although still quite high.

CHAPTER 3:

SPECTRAL ANALYSIS AND HIERARCHICAL CLUSTERING

This chapter summarizes the methods used in the analysis of the events in multiplet 18, especially the spectral analysis and hierarchical agglomerative clustering. R software was used for all calculations and is publicly available at www.R-project.org. Scripts for the analysis presented in this thesis are found in the Appendix A.

3.1 Spectral Density Estimation

This section uses the notation of Diggle (1991) and Venables and Ripley (1999) in summarizing the development of the periodogram as an estimate of the spectrum.

The model for the j^{th} event waveform $\{a_j(t), t = 1, 2, \dots, T\}$ is $\{A_j(t), t = 1, 2, \dots, T\}$, a set of regularly spaced, stationary Gaussian random variables with zero mean and autocovariance $\gamma_i(k) = \operatorname{Cov}\{A_j(t), A_j(t-k)\}$. The Fourier transform pair for the spectrum and autocovariance of A_j is

$$f_j(\omega) = \sum_{-\infty}^{\infty} \gamma_j(k) e^{i\omega(k)} \tag{3.1}$$

and

$$\gamma_j(k) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f_j(\omega) e^{ik\omega} \mathrm{d}\omega. \tag{3.2}$$

It can be shown by even symmetry of $\gamma_j(k)$ and trigonometric substitution that

$$f_j(\omega) = \gamma_j(0) + 2 \sum_{k=1}^{\infty} \gamma_i(k) \cos(k\omega) \tag{3.3}$$

and

$$\gamma_i(k) = \frac{1}{\pi} \int_0^{\pi} f_i(\omega) \cos(k\omega) \mathrm{d}\omega, \tag{3.4}$$

where $\gamma_j(0)$ is the variance of A_j.

For the j^{th} observed waveform, a_j, the sample autocovariance, $g_j(k)$, estimates $\gamma_j(k)$ and is written

$$g_j(k) = \frac{1}{n} \sum_{t=k+1}^{T} a_j(t) a_j(t-k) \quad \text{for integer } k \tag{3.5}$$

and $n = T - (k+1)$. The periodogram $I_j(\omega)$ then estimates $f_j(\omega)$ and is written

$$I_j(\omega) = g_j(0) + 2 \sum_{k=1}^{T-1} g_j(k) \cos(k\omega), \tag{3.6}$$

where $g_j(0)$ is the sample variance and $\omega \in (1, T/2)$ Hz.

Assuming a stationary Gaussian model for $A(t)$ ensures that the variability in

$A(t)$ can be wholly explained by $\gamma(k)$. Hence, the periodogram is also viewed as a partition of the total sum of squared errors into orthogonal components at the Fourier frequencies. This property results in nice statistical attributes for the periodogram ordinates (e.g., independence, χ^2 distributions), which allow for additional statistical analysis, if desired (Diggle, 1991). As a last step, I_j is normalized by $\max\{I_j\}$ so that $I_j \in [0, 1]$. These power spectra will be compared with hierarchical agglomerative clustering, which is described next.

3.2 Hierarchical Agglomerative Clustering

In agglomerative clustering, each event starts out as its own cluster, clusters are agglomerated (combined) into larger ones, culminating in a single cluster that contains all events. Agglomeration decisions are based on a dissimilarity metric, among which there are many to choose, including Euclidean distance, Mahalanobis distance, and Jeffrey's divergence (McLachlan, 2004). For the purpose of clustering microseismic events, I define spectral and amplitude distances as dissimilarity metrics.

Among clustering algorithms, hierarchical agglomerative clustering is attractive for several reasons: 1) unlike $k-$means (Hartigan, 1975), the number of clusters is not required in advance, 2) since it works on a distance matrix, it is robust to the order of events, and 3) interactive determination of the number of clusters provides flexibility, with the added advantage of quickly being able to identify outliers and singletons. Hierarchical agglomerative clustering is a widely used algorithm, successfully applied in a variety of disciplines, including seismic applications (Bardainne *et al.*, 2006; Rowe *et al.*, 2002).

Starting with the P-wave arrival, a_j is trimmed to the length of the shortest duration event so that the Fourier frequencies are constant between events and interpolation is not necessary. At the deepest station, the shortest length event is 1598 samples, which is then padded with zeros, so that $T = 1600$, to optimize the discrete Fourier transform. The result is 800 Fourier frequencies with a delta-frequency of 0.937 Hz. Squared spectral distance between events i and j is

$$d_{Freq(i,j)} = \left[I_i(\omega) - I_j(\omega)\right]^T \left[I_i(\omega) - I_j(\omega)\right] \tag{3.7}$$

with $I(\omega), \omega$ in the frequency band [1,375] Hz. Initially the clustering was performed using ω in [1,750] Hz. Event spectra contain little information or variability after approximately 350 Hz, and the results did not change when the higher frequencies were omitted, so discussion and results using fewer frequencies is presented.

A dendrogram is the analysis tool used to determine an appropriate number of clusters; an example is shown in Figure 3.1. The y-axis is "Dissimilarity," and horizontal lines show where agglomerations occur. Dissimilarity is the weighted distance (either spectral or temporal), as shown by (3.7) or (3.8) below.

Conceptually, all pairwise squared distances between events are calculated and put in a lower triangular matrix D with dimension n x n, where n is the number of events ($n(n-1)/2$ pairwise distances). The lowest horizontal line in Figure 3.1 is at $\min\{D\}$ and notes the first agglomeration. At the next step, D is recalculated, but its dimensions are $(n-1) \times (n-1)$, where the rows and columns corresponding to the clustered events are omitted and a new row calculating the distance between the cluster centroid and the other events is added. At this stage, there are several methods by which to choose the next agglomeration. Using $\min\{D\}$ throughout equates to a nearest neighbor approach, but does not necessarily result in the most compact clustering (Ward, 1963). Instead, Ward's minimum variance method is used (Ward, 1963), which minimizes the within-cluster variance (Var) at each agglomeration step, resulting in the most compact clusters possible. Rather then calculating $\mathrm{Var}[d_{Freq(i)}, d_{Freq(j)}]$ for each possible combination of clusters at each step, the Lance-Williams algorithm is used to recursively calculate dissimilarity (Lance and Williams, 1967). Here, for clusters C_i, C_j, and C_k with sizes n_1, n_2, and n_3, between-cluster dissimilarity after agglomerating clusters C_i and C_j is

$$d_{Freq(i,j)k} = \alpha_i d_{Freq(ik)} + \alpha_j d_{Freq(jk)} + \beta d_{Freq(ij)}, \tag{3.8}$$

where $\alpha_i = (n_i + n_3)/(n_1 + n_2 + n_3)$ and $\beta = (-n_3)/(n_1 + n_2 + n_3)$. The cluster with the minimum variance is the one that minimizes d_{Freq} over all i, j, k. Ward's method only works with true distance measures, which is a motivation for using Equation 3.7.

Time Domain

The clustering algorithm is identical in the time domain, but instead of $d_{Freq(i,j)}$, the measure is

$$d_{Time(i,j)} = [a_i(t) - a_j(t)]^T \left[a_i(t) - a_j(t)\right], \quad t = \{1, 2, \ldots T\}, \tag{3.9}$$

where $a_i(t)$ is the peak normalized amplitude at time t. Ward's minimum variance is used to make agglomeration decisions, just as in the frequency domain. Here, dissimilarity becomes

$$d_{Time(i,j)k} = \alpha_i d_{Time(ik)} + \alpha_j d_{Time(jk)} + \beta d_{Time(ij)}. \tag{3.10}$$

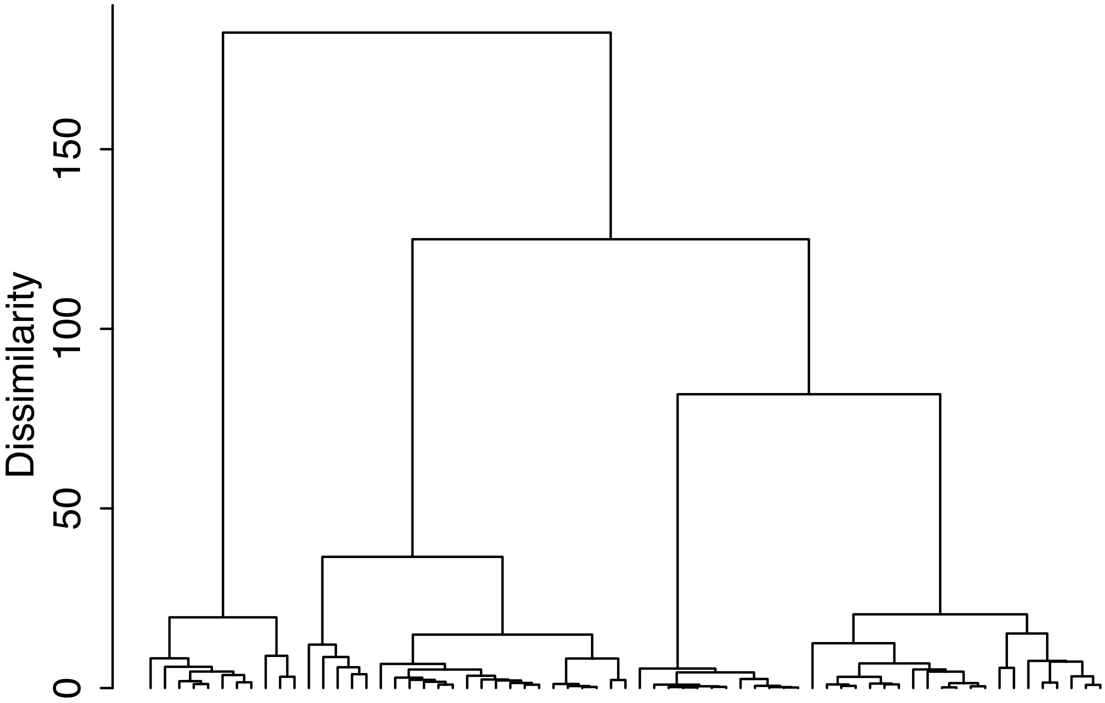


Figure 3.1: Example dendgrogram. The greater the dissimilarity, the later clusters will be agglomerated. In the frequency domain, dissimilarity has units (*intensity*)^2, in the time domain, (*amplitude*)^2.

CHAPTER 4:

CLUSTERING IN THE FREQUENCY DOMAIN

Based on a strong correlation of the wave fields in the time domain, events are determined to be of the same multiplet, located using a single master event. As seen in the lower panel in Figure 2.10, these events cluster into two areas that cross a velocity boundary, assuming the velocity model of Figure 2.7 is correct at the location of the events. However, even under this assumption, relative depth uncertainty vitiates velocity layer differences. Additional evidence that they represent different fractures, or a fracture that crosses a velocity boundary, lies in the spectral analysis.

The magenta dots in Figure 2.10, corresponding to multiplet 18 events, organize into western and eastern groupings and are named Groups 1 and 2, respectively. Of the 70 events in multiplet 18, 3 events do not have P-wave arrivals and were dropped from the analysis. The remaining multiplet 18 spectra, $\{I_i, i = 1, 2, \ldots, 67\}$, were clustered using methods of Chapter 3. The cluster dendrogram is shown in Figure 4.2, with color-coded rectangles corresponding to cluster membership. Height on the y-axis represents dissimilarity, and the higher horizontal agglomeration line indicates greater spectral dissimilarity. Group 2 I's (denoted $\{I_{G2}\}$) are agglomerated together early and only join Group 1 I's (denoted $\{I_{G1}\}$) at the final agglomeration step, indicating that $\{I_{G1}\}$ are markedly different than $\{I_{G2}\}$. Figure 4.1 plots the average spectrum of each group ($\bar{I}_{G1}, (n_{G1} = 56)$ and $\bar{I}_{G2}, (n_{G2} = 11)$); frequency content differs greatly between $\bar{I}_{G1}$ and $\bar{I}_{G2}$, corroborating evidence that these events originate in regions with different elastic properties.

It is reasonable to consider subdividing $\{I_{G1}\}$ into two subgroups based on Figure 4.2; the two subgroups in $\{I_{G1}\}$ are named Groups 1E and 1W (denoted $\{I_{1W}\}$ and $\{I_{1W}\}$, respectively) and their average spectra ($\bar{I}_{1E}$ and $\bar{I}_{1W}$) are shown in Figure 4.3. Although similar, the $\bar{I}_{1E}$ has more power in the higher frequencies than $\bar{I}_{1W}$.

Cross Correlation

The cross correlation of $\bar{I}_{1E}$ and $\bar{I}_{1W}$ is shown in Figure 4.4. Cross correlations are computed at discrete lags, which in the frequency domain correspond to discrete Fourier frequencies, thus the vertical alignment of lags in Figure 4.4, where each lag corresponds to a 0.93 Hz frequency shift. The maximum cross correlation of 0.82 at lag 1.87 Hz shows the strong similarity between $\{I_{1W}\}$ and $\{I_{1E}\}$ and the Group 1E frequency shift.

Similarly, $\{I_{G1}\}$ were cross correlated with both $\bar{I}_{1E}$ and $\bar{I}_{1W}$ in Figure 4.3. Maximum cross correlations and corresponding lags between $\{I_{G1}\}$ and $\bar{I}_{1W}$ are shown in Figure 4.5. As expected, the Group 1W events align at lag zero – the maximum correlation between $\{I_{1W}\}$ and $\bar{I}_{1W}$ is not shifted for any Group 1W event. The Group 1E events are all shifted positively with respect to Group 1W, indicating that $\{I_{1E}\}$ have slightly higher frequency content than $\bar{I}_{1W}$, and, in general, $\{I_{1W}\}$. While this difference in frequencies is only one or two lags (0.93 and 1.87 Hz), it has com-

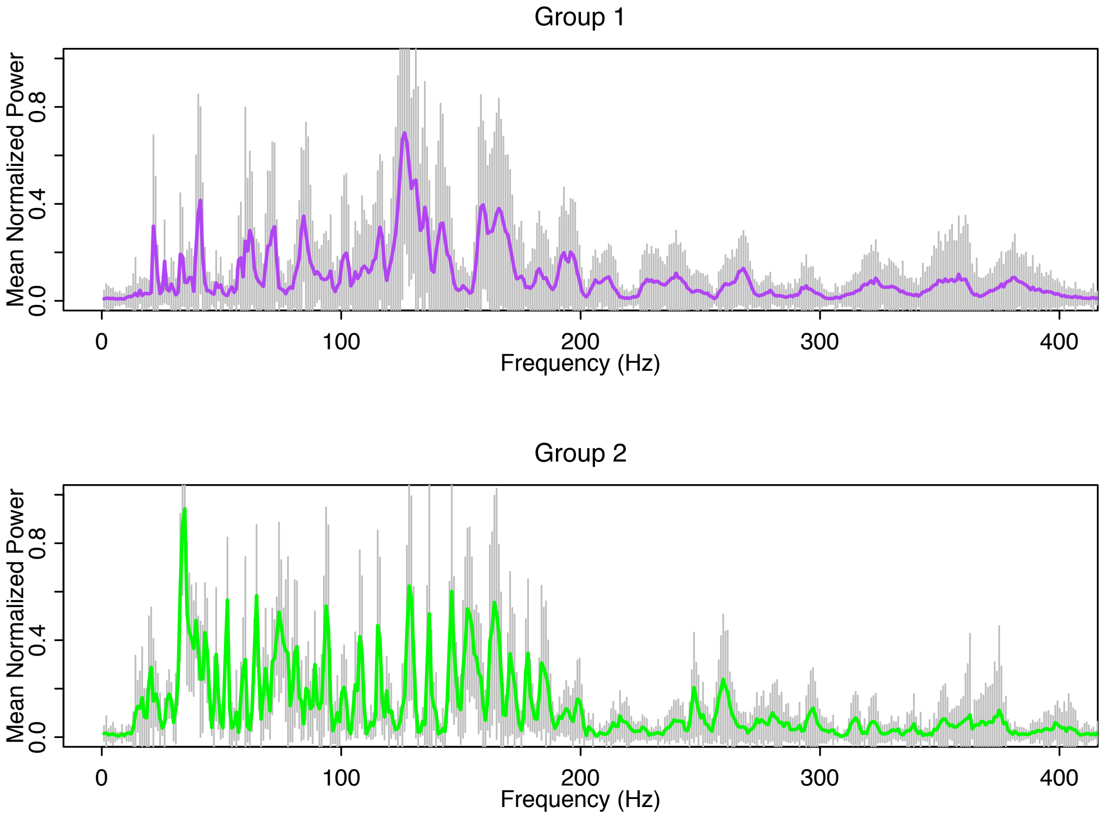


Figure 4.1: Average peak-normalized spectra, $\bar{I}_{G1}$ and $\bar{I}_{G2}$ for the western (G1) and eastern (G2) sets of multiplet 18 events. The colored line is the the mean by frequency whereas the light gray lines are individual 95% confidence intervals. Spectral energy is present generally between 20 and 200 Hz for both groups, but intensity differs.

pelling impact, as can be seen in Figure 4.6. The clusters organize spatially into two non-overlapping groups, indicating the spectra originate from two nearby but distinct spatial regions.

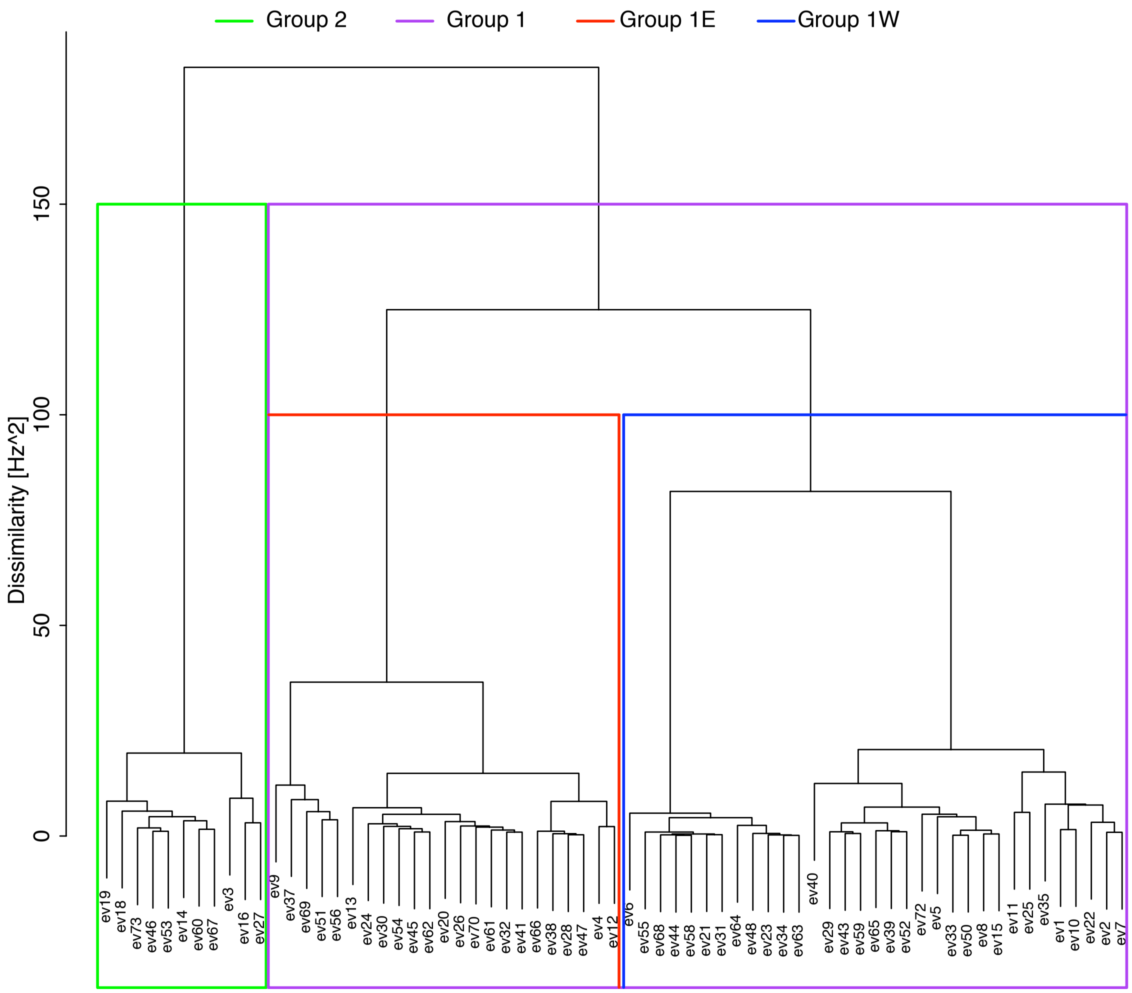


Figure 4.2: Dendrogram for $\{I\}$ with rectangles identifying cluster membership, with a first-order distinction between Group 1 and 2 annotated in Purple and Green, respectively. A secondary division in Group 1 in 1E and 1W is suggested in Red and Blue, respectively.

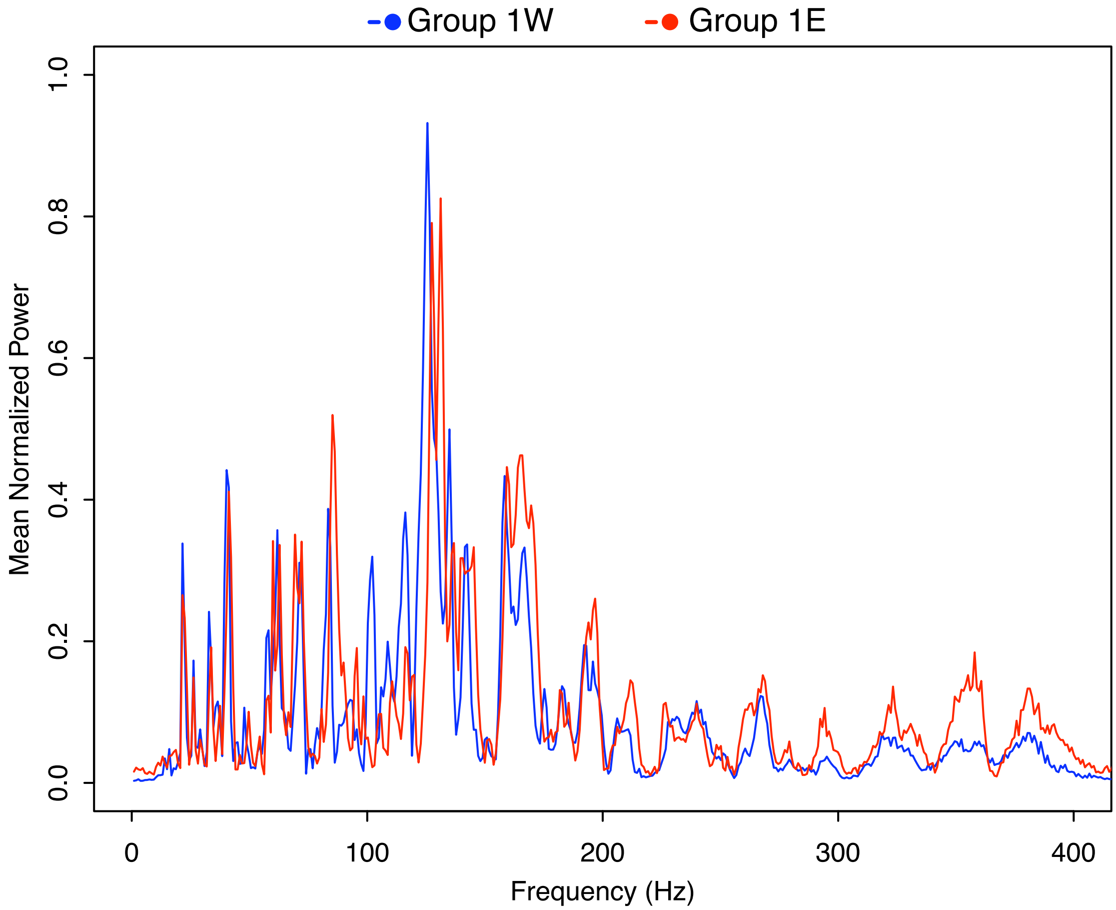


Figure 4.3: Mean spectra $\bar{I}_{1W}$ and $\bar{I}_{1E}$ appear to show a small but distinct shift in frequency.

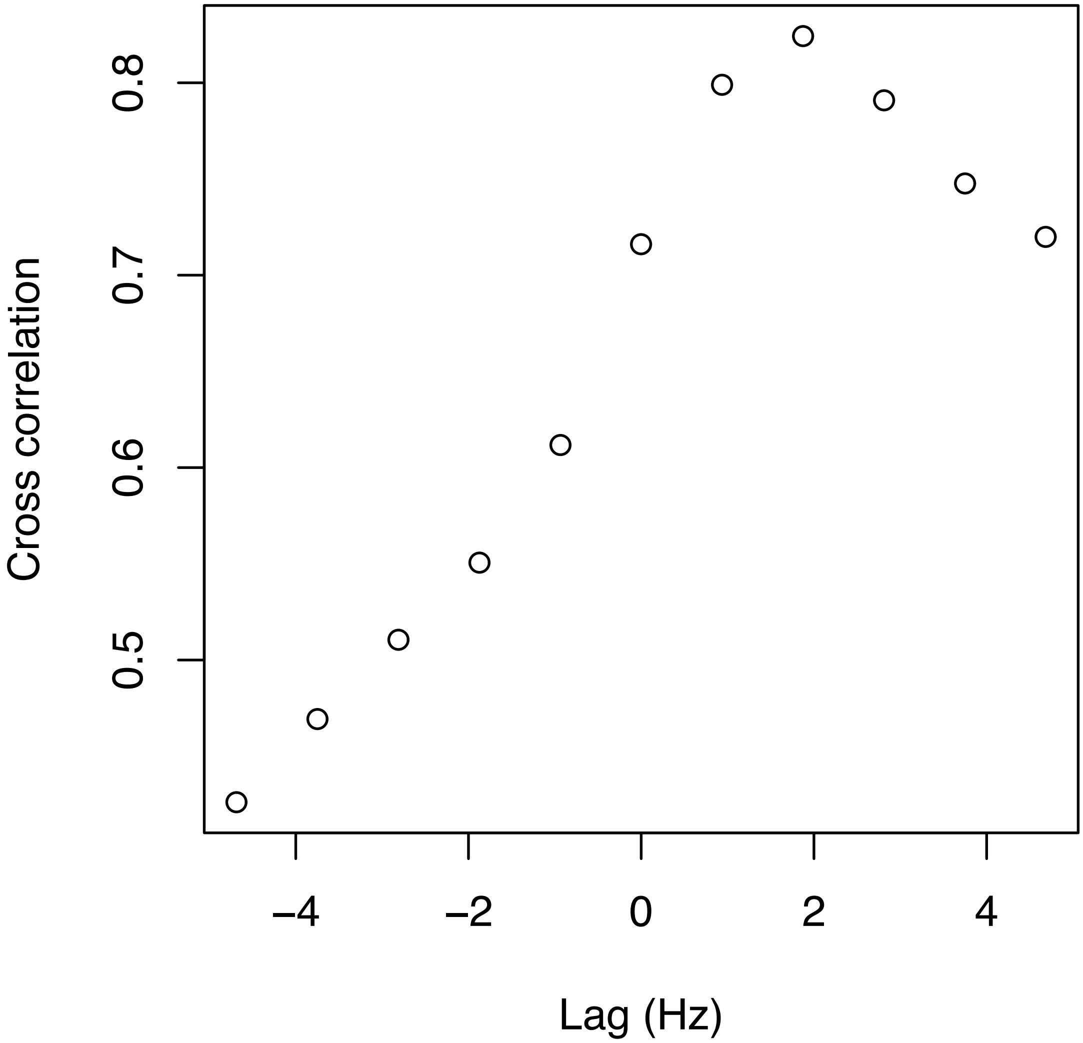


Figure 4.4: Cross correlation coefficients for $\bar{I}_{1W}$ cross correlated with $\bar{I}_{1E}$. The maximum at a positive lag of 2 Hz indicates a positive frequency shift of $\bar{I}_{1W}$ with respect to $\bar{I}_{1E}$, as alluded to in Figure 4.3.

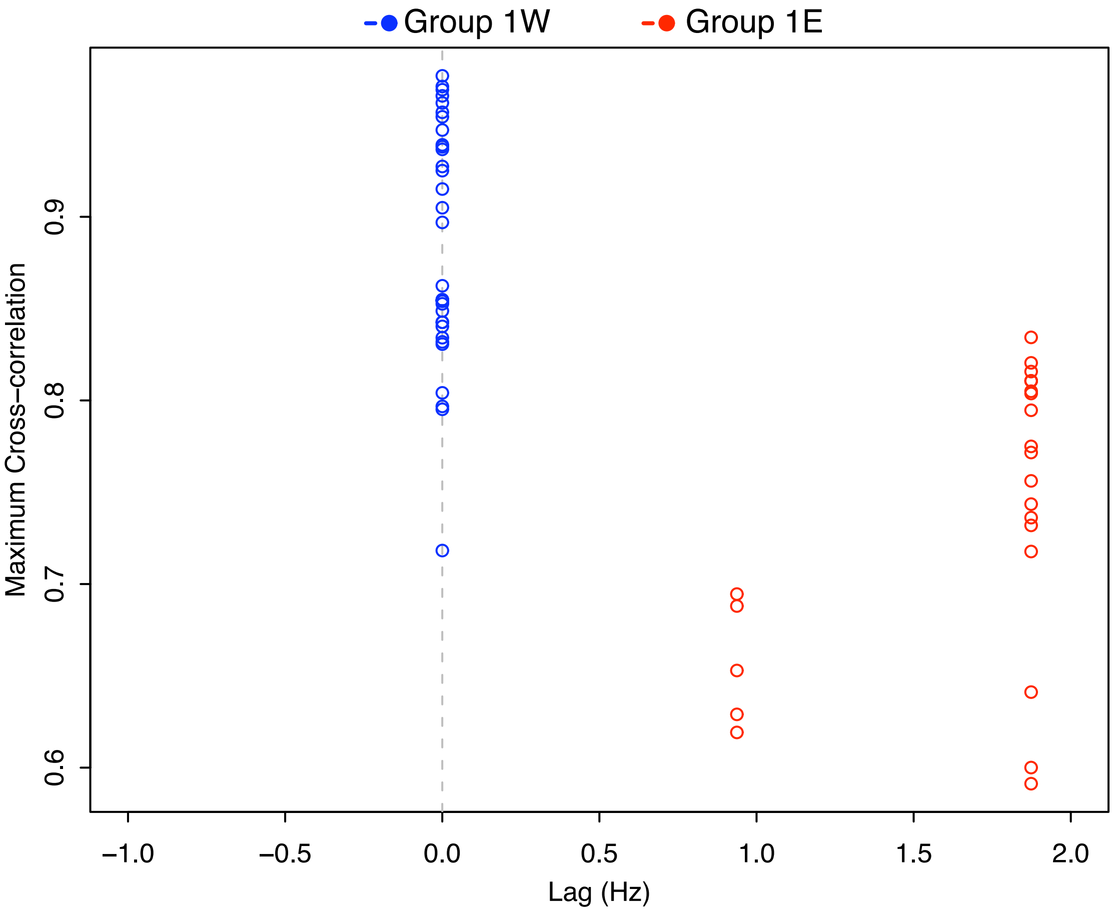


Figure 4.5: $\{I_{1W}\}$ (blue) and $\{I_{1E}\}$ (red), cross correlated with $\bar{I}_{1W}$ show little variance, indicating that correlations of the individual events are robust to the mean. The 2 Hz shift noted in Figure 4.4 is also evident here.

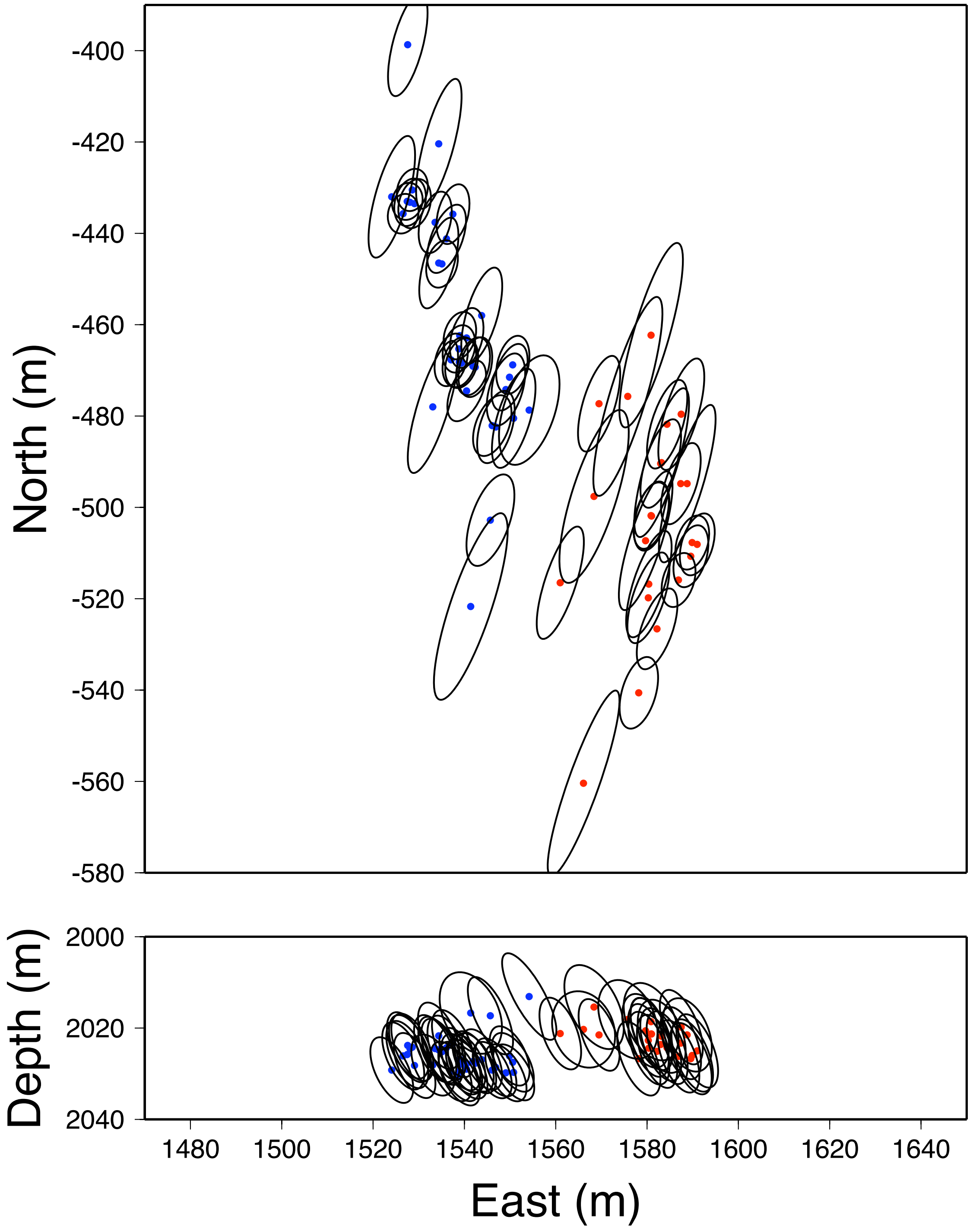


Figure 4.6: $\{I_{G1}\}$ locations, color coded by spectral membership: $\{I_{1W}\}$ (blue), $\{I_{1E}\}$ (red), show that 1W and 1E organize clearly in space. Black ovals represent location 1σ uncertainty ellipses.

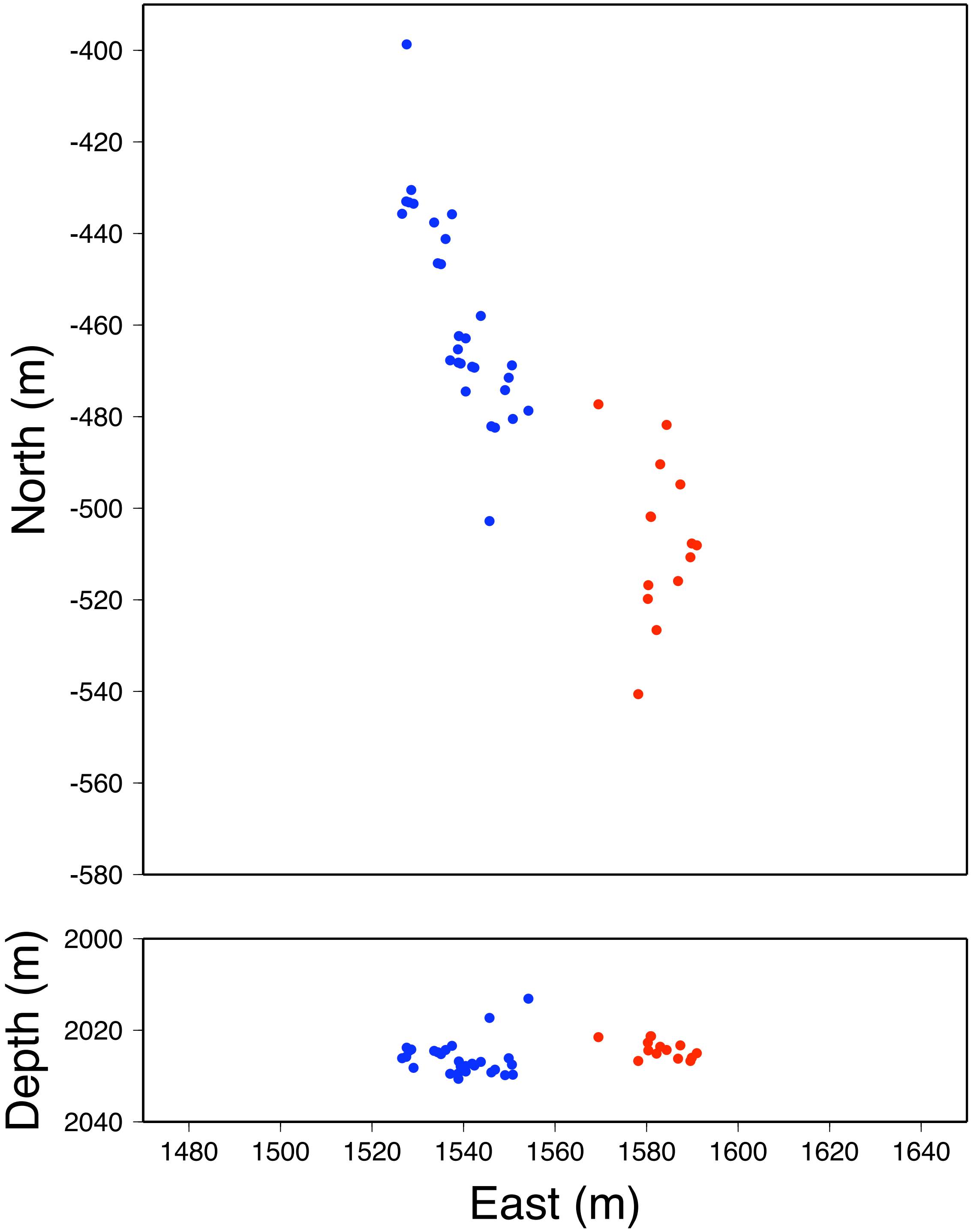


Figure 4.7: The 80% of $\{I_{G1}\}$ with the smallest location uncertainty, as calculated by the area of uncertainty ellipses. These events organize into two non-overlapping, parallel regions, ($\{I_{1W}\}$ (blue) and $\{I_{1E}\}$ (red)) indicative of en echelon fracturing.

4.1 Discussion

Aside from the spectral differences that are indicated by color-coding in Figure 4.6, neither event locations nor waveforms exhibit particularly distinguishing features that differentiate events within $\{I_{G1}\}$. Size and orientation of relative error ellipses blurs the distinction between sub-groups and, along with high time-domain waveform correlation, makes arguing subsurface structure difficult. Spectral correlation, however, reveals distinguishing characteristics that hint at subsurface structure. The spectral clusters organize spatially into two non-overlapping groups, indicating two nearby but distinct spatial regions. Structure is further identified by omitting events with the largest 20% of error-area, as seen in Figure 4.7.

If attenuation was a factor in the spectral separation of the events in $\{I_{G1}\}$, one would expect the opposite of what is observed: events farther from the receiver, $\{I_{1E}\}$, would have less spectral energy in the higher frequencies than $\{I_{1W}\}$. Instead, we attribute small spectral differences between events in $\{I_{G1}\}$ to either two closely separated fractures, or one fracture transitioning (laterally) into a material with varying elastic properties, or different mechanical behavior of individual fractures.

In a temporal analysis of events, Rutledge and Soma (2010) observe that higher magnitude events occur near time gaps in seismicity, indicating that the structure is composed of discrete segments, such as en echelon jogs where stress would tend to concentrate near discontinuities. Additionally, a magnitude distribution analysis shows a population of earthquakes dominated by small events, which is generally considered an indication of many small discontinuous faults accommodating strain accumulation. However, the single- well array geometry and large source-receiver distances prevent resolving detailed subsurface structure in the time domain.

The results of spectral clustering add insight to our understanding of fracturing at Aneth Field. In contrast to the striking spectral differences between $\{I_{G1}\}$ and $\{I_{G2}\}$, which are hypothesized to originate from different velocity layers or have different fracture orientations, the subtly different spectra of $\{I_{1W}\}$ and $\{I_{1E}\}$ indicate some other cause. Their spectral power differences are small, but accumulate in the statistical clustering algorithm, suggesting that these sub groups represent different fractures within the same medium. The highest quality (smallest uncertainty) events from $\{I_{G1}\}$ are shown in Figure 4.7, color coded by spectral group. These events suggest separate, parallel fractures that are indicative of en echelon fracturing, supporting the conclusions of Rutledge and Soma (2010).

Additionally, the fracture pattern observed is consistent with both the microseismic event lineament shown in Figure 2.6 as well as the overall structure of the basin. The lineament extends NW-SE, similar to the surface folds shown if Figure 2.1 and surface deformation band patterns indicated by Figure 2.3. The fractures observed are oblique to the lineament, a characteristic of echelon fracturing. Finally, en echelon fractures are planar structures. Although the formation depths below the oil producing zones are not well understood in the region of microseismicity, assuming the layer thickness contour plot in Figure 2.4 represents thicknesses patterns of the underlying formations, the consistent depths of multiplet 18 events indicates they occur within the same structure, and are therefore planar as well.

As discussed in Zhang and Thurber (2003), event epicenters estimated from double difference algorithms may be biased if inter-event distances exceed the scale length of the velocity model. An assumption is that events are from the same fracture or medium. This is particularly relevant in highly heterogeneous sites such as the Aneth field. While shallower stations indicate that both $\{I_{G1}\}$ and $\{I_{G2}\}$ events belong to the same multiplet, spectral analysis of likely guided waves at the deepest station shows otherwise and suggests that these events should be separated in a refined location analysis.

CHAPTER 5:

CLUSTERING IN THE TIME DOMAIN

In Figure 2.11, the bottom five traces are from Group 2. One can visually differentiate these waveforms from the top 20, which are randomly selected from Group 1, and include events from both Group 1W and Group 1E. Their subtle spectral differences are virtually impossible to visually detect in the time domain. Hierarchical agglomerative clustering was applied to multiplet 18 events in the time domain to investigate whether spectral differences are detected using a parallel algorithm.

As in the frequency domain, events are truncated to the shortest length event and then are peak normalized. Figure 5.1 shows the dendrogram from time domain clustering. Although dissimilarity is unitless, it is not comparable between the domains. As discussed in Chapter 3, d_{Freq} is the sum of 400 values, whereas in the time domain, d_{Time} is the sum of 1597 values. Dissimilarities $d_{Freq(ij)k}$ and $d_{Time(ij)k}$ are not directly comparable because spectral differences are manifested in the time domain as both amplitude and waveform shape differences. Generally speaking, then, for fixed i, j, k, $d_{Freq(ij)k} < d_{Time(ij)k}$. To illustrate this, Figure 5.2 shows spectra and filtered waveforms for two events. Events show highly similar spectra, especially in the region $(20, 50)$ Hz, however, when bandpass filtered over $(20, 50)$ Hz, waveforms show both amplitude and wave shape differences.

From Figure 5.1, three clusters seems reasonable. However, if we consider three clusters without first considering two, we miss the important result that events do not have the same spatial clustering as in the frequency domain. Figure 5.3 shows the spatial results of two clusters in the time domain. The far southeast events are not a cluster by themselves, as in the frequency domain (recall the western and eastern groups in Figure 2.6 being hierarchically clustered into Groups 1 and 2 in Figure 4.2), but are joined with events in Group 1E. Analogous to the method used in the frequency domain, Figure 5.4 shows cross correlations between $\{a\}$ and $\bar{a}_{G1}$ where $G1$ membership is determined in the time domain (Figure 5.1). Group 2 events show generally low correlations at negative lags, indicating a negative time shift off the $\bar{a}_{G1}$. A considerable number of Group 1 events have maximum correlations at lag 0, but there is significant variation in both the correlations and lags within this group.

5.1 Discussion

Intuitively, the time and frequency domains should give similar results, because, as noted in Zadler *et al.* (2005), the domains contain the same information. However, a few observations are worth considering.

In the frequency domain, a peak in a power spectrum indicates energy present across the duration of the signal in the time domain. In this sense, each frequency in the spectrum represents the entire time series. Frequency domain clustering, then, compares multiple characteristics of the entire time domain. Conversely, dissimilarity in the time domain compares individual observations within signals. Along those lines, the P wave pick has a larger influence on the results in the time domain than

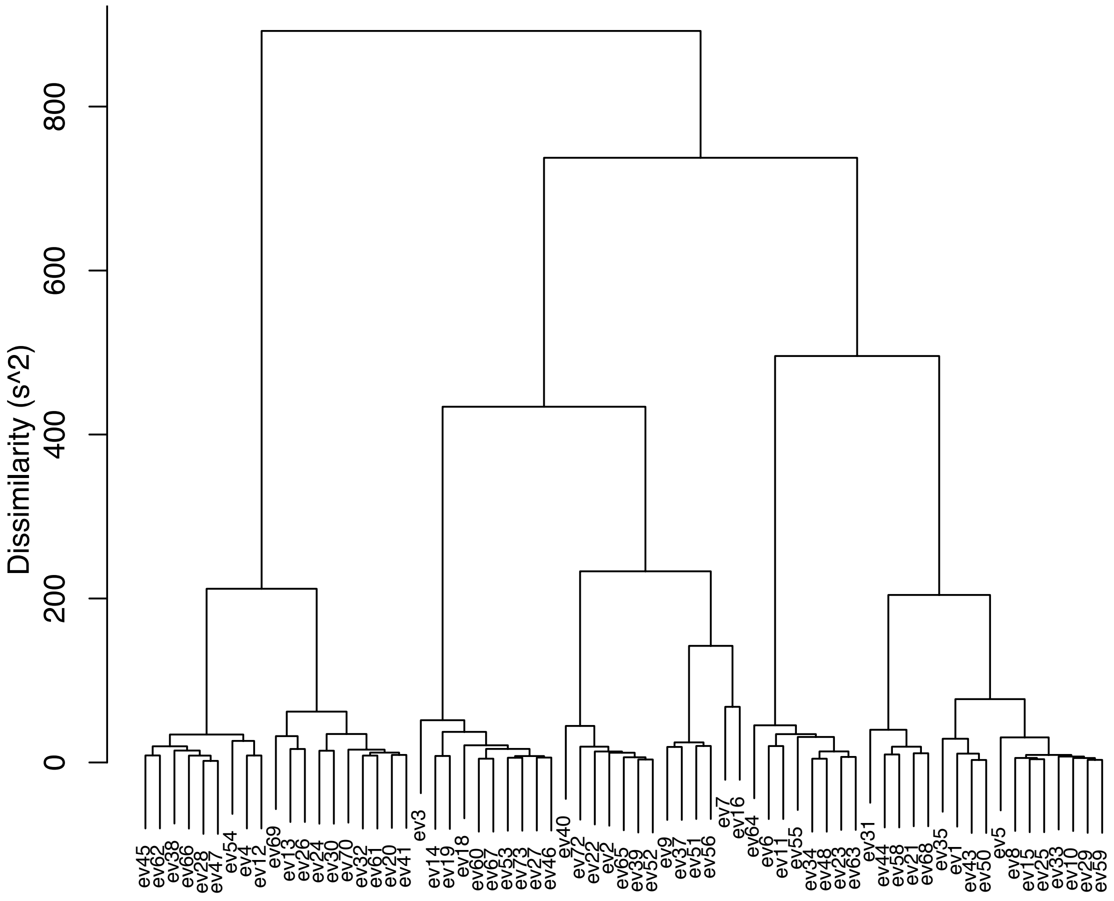
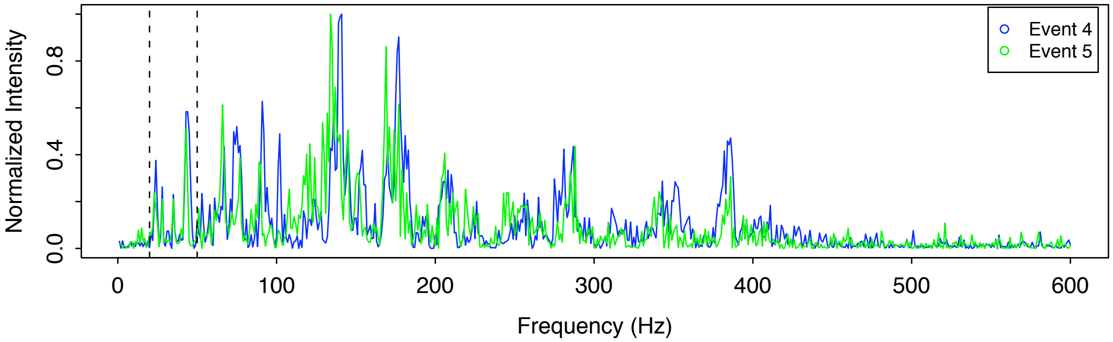
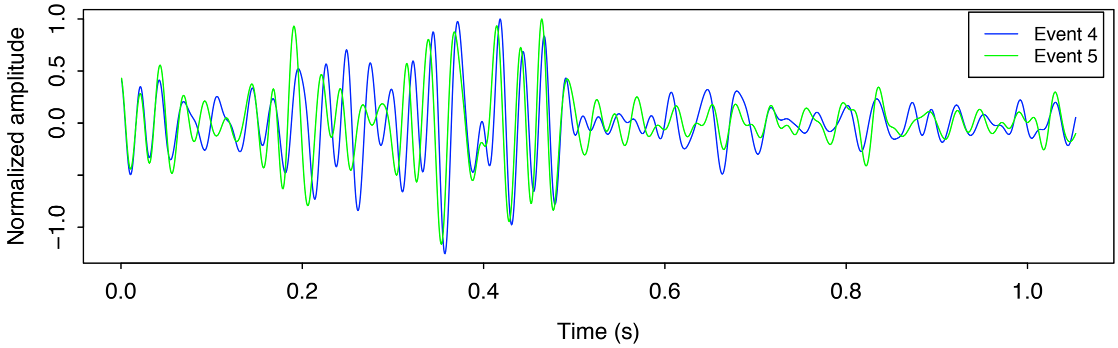


Figure 5.1: Time domain dendrogram for multiplet 18 events, which differs in cluster membership from Figure 4.2.



(a) normalized spectra



(b) normalized filtered waveforms

Figure 5.2: Events 4 and 5 spectra (I_4, I_5) and filtered waveforms. Event spectra are nearly identical, especially in the region $[20, 50]$ Hz (between the dashed vertical lines). Bandpass filtered waveforms on $[20, 50]$ Hz show peak shape and amplitude dissimilarities.

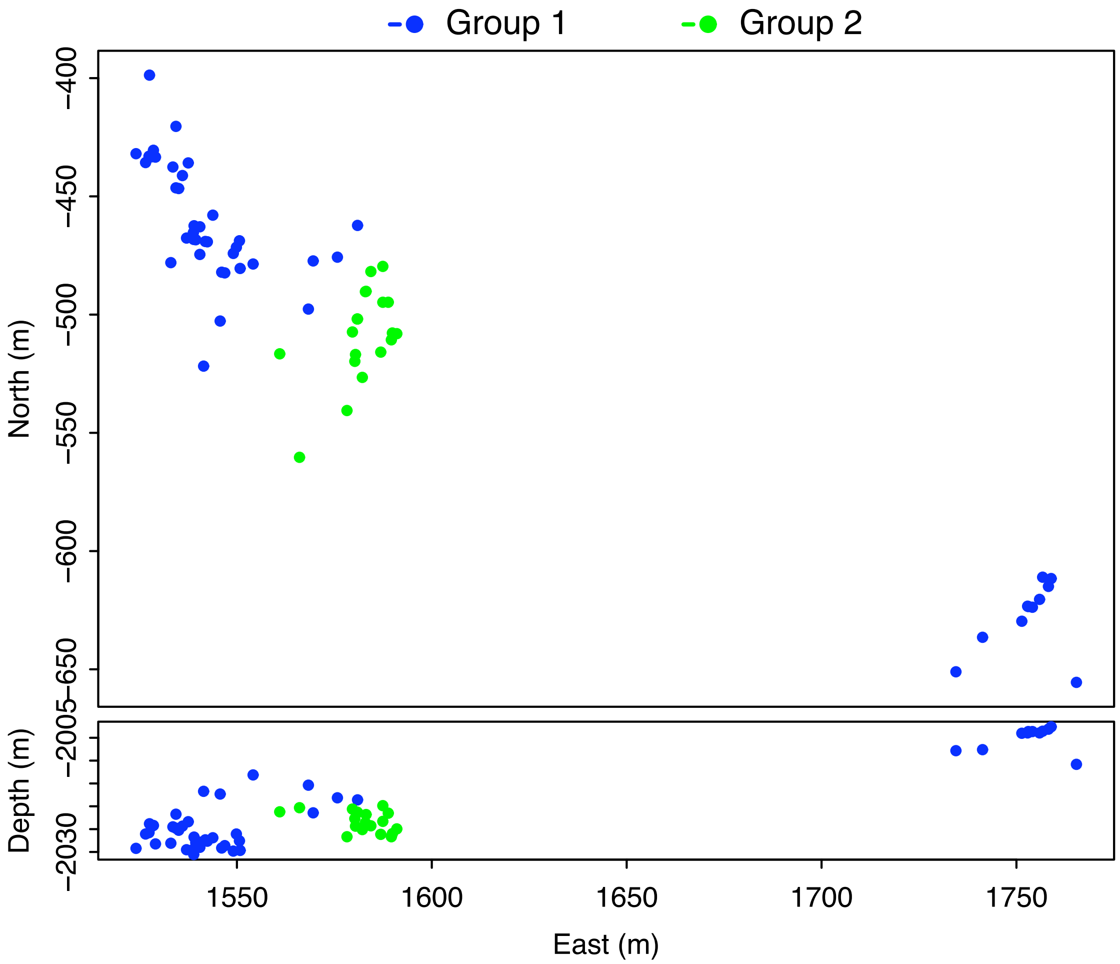


Figure 5.3: Multiplet 18 event locations color coded by time domain cluster membership, which differs from frequency domain results $\{I_{G1}\}$ and $\{I_{G2}\}$.

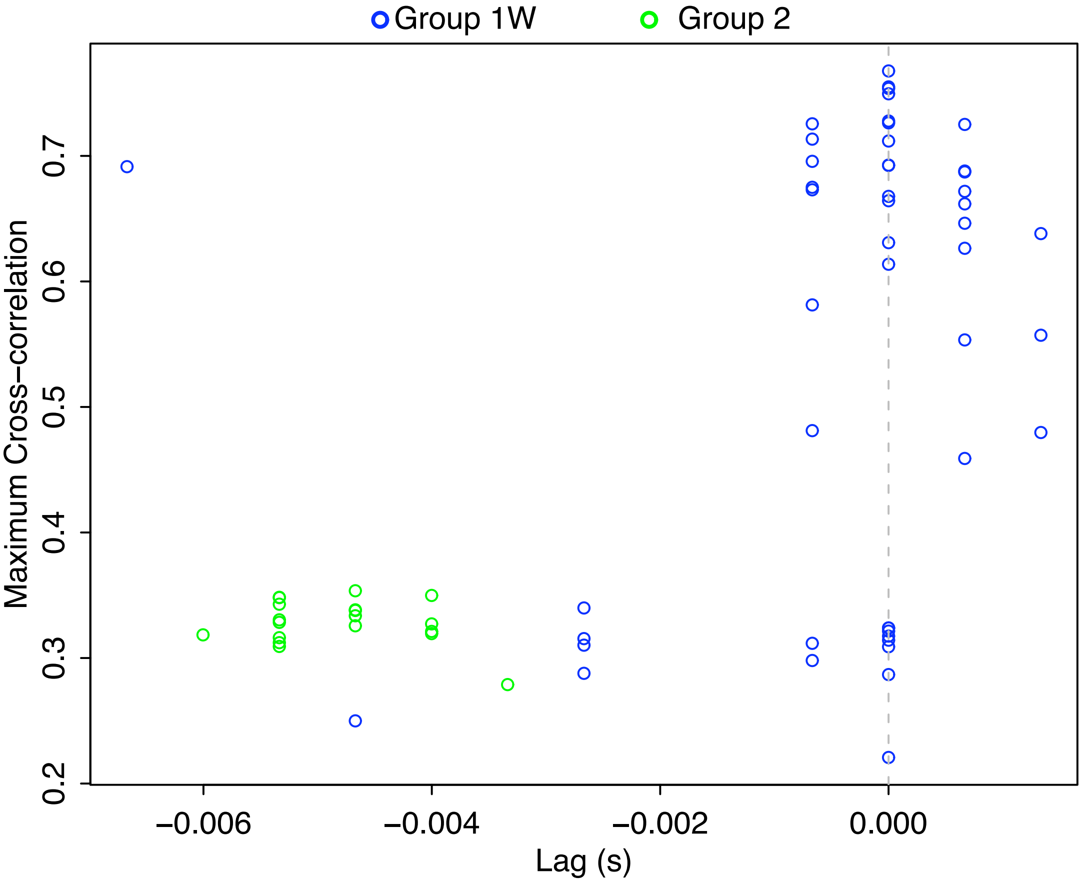


Figure 5.4: Maximum correlation between $\{a\}$ and $\bar{a}_{G1}$ color coded by time domain group membership. Correlation coefficients are generally lower in the time domain and show more variability about the mean.

in the frequency domain. If two signals are identical but for the P pick, a constant time shift in the series results in a large dissimilarity value in the time domain but a very small one in the frequency domain.

Additionally, frequency domain clustering as performed here does not incorporate the phase spectrum, whereas in the time domain, both frequency and phase are being compared simultaneously. As with the influence of the P pick, a simple phase shift could cause a large dissimilarity value in the time domain, altering the cluster results between the two domains.

Because P picks are based on waveform correlations, the effects of P pick and phase shifts are minimized to the extent possible but may become more of an issue if comparing events from different multiplets. However, there are still differences between the two domains that cannot be explained so far. The events cluster in a completely different order in the time domain, resulting in different cluster membership for two clusters.

While the goal of the time domain clustering was to determine whether similar spatial organization resulted using the same algorithm, the amount of time spent studying and interpreting the results was significantly less, and the differences between the two results deserves more attention.

CHAPTER 6:

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

The events analyzed here occur between 2002 and 2030 m and were grouped into a single multiplet based on their waveform correlations at stations between 800 and 1550 m. Geophones between 1590 and 1705 m were not used in the original multiplet and location analysis because of their weak arrivals and the likelihood of guided waves. The deepest geophone, at 1705 m, is the receiver closest to the events, and waves traveling between events and this receiver are the least affected by attenuation and geometrical spreading. Additionally, if, waves at this geophone are head waves, their direct waves traverse layers beneath the events, rather than only layers above. This thesis demonstrates that information regarding location, and potentially structure, may be gleaned from this geophone.

Visual inspection of waveforms in the time domain indicate two subsets of events (Groups 1 and 2). Power spectra between these groups are markedly different, leading to the supposition that events may be located in two different lithologic units or possibly have different fracture orientation. Spectral clustering further categorizes events within Group 1 that may be related to separate fractures. This fracture structure is not well-identified in the time domain, likely due to spectral differences that manifest in the time domain as both amplitude and wave shape differences. Removing events with the largest uncertainty ellipses suggests that this method may be successful at identifying individual en echelon fractures, thereby giving an indication of shear zones and non-coaxial shear fracturing.

6.2 Future Work

The results presented here include one component from a single station. There are several obvious avenues for future work, both to further explore the technique and to utilize its results.

Possible future work using the Aneth data set includes analysis of additional components and stations. Preliminary frequency domain analysis at Geophone 2, the second-deepest receiver ($\sim$ 1689 m), yields ambiguous results, which we postulate may be the result of geophones being located at lithologic transitions (see Figure 2.7). Frequency domain comparisons between the additional deep geophones might provide further illumination. Additionally, comparing location uncertainties before and after spectral clustering would give an indication of its ability to impact location analysis.

Zadler *et al.* (2005) illustrate the sensitivities of time and frequency domain analysis with respect to coda wave interferometry and spectroscopy and discuss the ability of each to reveal medium velocity changes. Similarly, as mentioned in Chapter 5.1, a better characterization of the difference between time and frequency clustering results would be useful, potentially increasing understanding of how nuances of spectral clustering are reflected in the time domain, as well as to possibly understand the geophysical properties that are reflected in the clustering in each domain.

Events in Figure 2.10 trend NW-SE. Analyzing additional multiplets would not only verify its usefulness at this site, but would also give a new cross-sectional look at the heterogeneity of the layers underlying the reservoir, which to date have not been greatly studied.

Spectral clustering in the frequency domain should be attempted at different sites to determine its broader usefulness and reproducability. Assuming reasonable utility, work remains to determine how to best incorporate results of spectral clustering into geophysical modeling.

Popular techniques to refine hypocenter distributions include *hypodd* (Waldhauser and Ellsworth, 2000) and *tomodd* (Zhang and Thurber, 2003), based on the double differences between P- and S-wave arrivals of closely spaced events. The methods rely on inter-event distances not to exceed the scale length of the velocity model, or bias may be introduced. This work suggests that spectral clustering may help predefine subsets of hypocenters, which in turn can then be relocated via double-difference techniques.

REFERENCES

Bardainne, T., Gaillot, P., Dubos-Sallée, N., Blanco, J., and Sénéchal, G. 2006. Characterization of seismic waveforms and classification of seismic events using chirplet atomic decomposition. Example from the Lacq gas field (Western Pyrenees, France). *Geophysical Journal International*, **166**(2), 699–718.

Carney, Stephanie. 2010. *Chapter 3, Suburface Structural and Thickness Mapping.* Tech. rept. Utah Geological Survey, Salt Lake City, UT.

Chidsey, Jr., Thomas C. 2009. *Surface and subsurface geological characterization of the Aneth Unit, Greater Aneth field, Paradox Basin, Utah - final report.* Tech. rept. Utah Geological Survey, Salt Lake City, UT.

Chidsey, Jr., Thomas C. 2010. *Chapter 1, Introduction of Surface and Subsurface Geological Characterization of the Aneth Unit, Greater Aneth Field, Paradox Basin, Utah.* Tech. rept. Utah Geological Survey, Salt Lake City, UT.

Christiansen, L. B., Hurwitz, S., Saar, M. O., Ingebritsen, S. E., and Hsieh, P. A. 2005. Season seismicity at western United States volcanic centers. *Earth and Planetary Science Letters*, **240**, 307 – 321.

Diggle, P. J. 1991. *Time Series: A Biostatistical Introduction.* Clarendon Press, Oxford.

Haney, M. M., van Wijk, K., Preston, L. A., and Aldridge, D. F. 2009. Observation and modeling of source effects in coda wave interferometry at Pavlof volcano. *The Leading Edge*, 554 – 560.

Hansen, S. E., Schwartz, S. Y., DeShon, H. R., and González, V. 2006. Earthquake relocation and focal mechanism determination using waveform cross correlation, Nicoya Peninsula, Costa Rica. *Bulletin of the Seismological Society of America*, **96**(3), 1003 – 1011.

Harris, D. B. 1991. A waveform correlation method for identifying quarry explosions. *Bulletin of the Seismological Society of America*, **81**(6), 2395 – 2418.

Hartigan, J. A. 1975. *Clustering Algorithms.* John Wiley & Sons, New York.

Hintze, L. F., and Kowallis, B. J. 2009. *Geologic history of Utah: Brigham Young University Geology Studies Special Publication 9.* Tech. rept. Brigham Young University, Salt Lake City, UT.

Lance, G. N., and Williams, W T. 1967. A general theory of classificatory sorting strategies: I. Hierarchical Systems. *Computer Journal*, **9**, 373–380.

Lay, T., and Wallace, T. C. 1995. *Modern Global Seismology.* Academic Press, San Diego.

Maxwell, S. C., and Urbancic, Th. I. 2001. The role of microseismic monitoring in the instrumented oil field. *The Leading Edge*, **20**(6), 636 – 639.

McLachlan, G. J. 2004. *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley & Sons, New Jersey.

Morgan, C. D., Carney, S., and Berg, M. Vanden. 2010. *Chapter 2, Surface Geology.* Tech. rept. Utah Geological Survey, Salt Lake City, UT.

Payan, C., Garnier, V., Moysan, J., and Johnson, P. A. 2009. Determination of third order elastic constants in a complex solid applying coda wave interferometry. *Applied Physics Letters*, **94**, 011904.

Phillips, W. S., Fairbanks, T. D., Rutledge, J T., and Anderson, D. W. 1989. Induced microearthquake patterns and oil-producing fracture systems in the Austin Chalk. *Tectonophysics*, **289**, 153 – 169.

Rowe, C. A., Aster, R. C., Borchers, B., and Young, C. J. 2002. An automatic, adaptive algorithm for refining phase picks in large seismic data sets. *Bulletin of the Seismological Society of America*, **92**, 1660 – 1674.

Rutledge, J., and Soma, N. 2010. *Chapter 12, Microseismic Monitoring of CO_2 Enhanced Oil Recovery in the Aneth Field.* Tech. rept. LA-UR-10-08213. Los Alamos National Laboratory, Los Alamos, NM.

Rutledge, J. T., and Phillips, W. S. 2003. Hydraulic stimulations of natural fracture as revealed by induced microearthquakes, Carthage Cotton Valley gas field, east Texas. *Geophysics*, **68**, 441 – 452.

Saar, M. O., and Manga, M. 2003. Seismicity induced by seasonal groundwater recharge at Mt. Hood, Oregon. *Earth and Planetary Science Letters*, **214**, 605 – 618.

Schaff, D. P., and Beroza, G. C. 2004. Coseismic and postseismic velocity changes measured by repeating earthquakes. *Journal of Geophysical Research*, **109**, 14. doi:200410.1029/2004JB003011.

Schaff, D. P., and Richards, P. G. 2004. Repeating seismic events in China. *Science*, **303**(5661), 1176–1178. doi:10.1126/science.1093422.

Schaff, D. P., and Richards, P. G. 2011. On finding and using repeating seismic events in and near China,. *Journal of Geophysical Research*, **116**. doi:201110.1029/2010JB007895.

Shumway, R. H. 2003. Time-frequency clustering and discriminant analysis. *Statistics & Probability Letters*, **63**, 307 – 314.

Shumway, R. H., and Stoffer, D. S. 2006. *Time Series Analysis and Its Applications: With R Examples*. 2nd edn. Springer-Verlag New York, LLC.

Snieder, R., and Hagerty, M. 2004. Monitoring change in volcanic interiors using coda wave interferometry: Application to Arenal Volcano, Costa Rica. *Geophysical Research Letters*, **31**, L09608.

Snieder, R., and Vrijlandt, M. 2005. Constraining relative source locations with coda wave interferometry: Theory and application to earthquake doublets in the Hayward Fault, California. *Journal of Geophysical Research*, **110**, B04301.

Song, F., Kuleli, H. S., Toksöz, M. N., Ay, E., and Zhang, H. 2010. An improved method for hydrofracture-induced microseismic event detection and phase picking. *Geophysics*, **75**(6), A74 – A52.

Streit, J. E., Siggins, A. F., and Evans, B. J. 2005. Predicting and monitoring geomechanical effects of CO_2 injection. *Pages 751 – 770 of:* Benson, Sally M. (ed), *Carbon Dioxide Capture for Storage in Deep Geologic Formations, Volume 2.* Elsevier Ltd.

US EPA Region IX UIC Program. 2007 (December). *Final Permit, Class III Water Injection Well, Permit No. NN207000002, Well Name: Aneth Unit C-113 LDVL SWD, San Juan County, Utah, Navajo Nation, API No. 4303731852.*

Venables, W. N., and Ripley, B. D. 1999. *Modern and Applied Statistics with S-PLUS.* 3rd edn. Springer-Verlag New York, LLC.

Waldhauser, F., and Ellsworth, W. L. 2000. A Double-Difference Earthquake Location Algorithm: Method and Application to the Northern Hayward Fault, California. *Bulletin of the Seismological Society of America*, **90**(6), 1353–1368.

Ward, J. H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**(301), 236–244.

Zadler, B. J., Grêt, A., and Scales, J. A. 2005. Spectroscopy versus interferometry: Resolving small changes. *Am. J. Phys.*, **73**(9), 837 – 834.

Zhang, H., and Thurber, C. H. 2003. Double-Difference Tomography: The Method and Its Application to the Hayward Fault, California. *Bulletin of the Seismological Society of America*, **93**(5), 1875 – 1889.

APPENDIX A:

R SCRIPTS

This appendix contains the R-scripts for this analysis, which was performed using Version 2.9.2, GUI 1.29, Tiger build, 32-bit. R software is platform independent and available publicly at www.R-project.org. Useful libraries, also available at the same website, include ”Rsac”, for loading SAC files and performing seismic analysis, and ”cluster” for statistical clustering algorithms.

```
#2012/06/27
################################################


#attach libraries
library(Rsac)
library(cluster)


#set paths
mymachine="dt"  #"lt" for laptop,  "dt" for desktop
if(mymachine=="lt"){
```

```
  #input paths
  toppath="/Users/dfagan/School/EOR/"
  myendian='big'
} else{
  toppath="/Users/deb/SchlumbergerEOR/"
}
datadir=paste(toppath,"Data/cluster_00018/",sep="")
codepath=paste(toppath,"RCode/",sep="")
figpath=paste(toppath,"Figures20110203/",sep="")
#################################################
#read in sac data
#get list of event directories
mydirs=list.files(topdir,pattern=glob2rx("d*"))

#get list of -component files in each event directory
fils=vector('list',length(mydirs))
for(i in 1:length(mydirs)){
  mypath=paste(topdir,mydirs[i],sep="")
  fils[[i]]=list.files(mypath,pattern=glob2rx("*E"))
}

#number of events
nev=length(mydirs)
```

```
#read in data by event
ev=vector('list',nev)
names(ev)=paste('ev',1:nev,sep="")
for(k in 1:nev){
  cat('k=',k,'; ')
  fnames=paste(topdir,mydirs[k],'/',fils[[k]],sep="")
  ev[[k]]=rsac(fnames,endian='big')
  names(ev[[k]])=paste('sta',1:length(ev[[k]]),sep="")
}
dt=ev$ev1$sta1$dt   #dt is the same for all events/stations
#################################################

#specify date of regional earthquake
eqday=158  #June 8, 2008
#################################################

#read in location data for events
  #convert feet to meters
  #remove events that do not have locations
evlocs=read.table(paste(toppath,"Data/AllEventLocs.txt",sep=""),
              header=T,as.is=T)
#convert feet to meters in event locs
ftm=0.3048
evlocs$north.m=evlocs$north.ft*ftm
```

```
evlocs$east.m=evlocs$east.ft*ftm
evlocs$depth.m=evlocs$depth*ftm

mymatch=match(mydirs,evlocs$event)

#check for unlocated events
evnoloc=mydirs[is.na(mymatch)]

#remove unlocated events
indx=which(is.na(mymatch),arr.ind=T)
ev=ev[-indx]

#############################################################
#read in multiplet 18 locations separately
c18locs=read.table(paste(toppath,'/Data/C18locs.txt',sep=""),
                header=T,as.is=T)
#convert feet to meters
ftm=0.3048
c18locs$north.m=c18locs$north.ft*ftm
c18locs$east.m=c18locs$east.ft*ftm
c18locs$depth.m=c18locs$depth*ftm

#####################################################
#read in geophone locations
```

```
staloc=read.table(paste(toppath,"Data/geophoneLocs.txt",sep=""),
                                 skip=4)
names(staloc)=c('sonde','md','tvd','north','east')


#convert feet to meters
ftm=0.3048
staloc[,c('north','east','tvd','md')]=
           ftm*staloc[,c('north','east','tvd','md')]


#####################################################
#source functions
calcCCF.r=function(ev1=sta11f1$ev1$amp,ev2=sta11f1$ev8$amp){
  #requires dt, tw
  #ev1 is reference event
  len=min(length(ev1),length(ev2))
  nwin=floor(len*dt/tw)
  win=tw/dt
  seq1=seq(1,len,floor(tw/dt))
  indx=rep(seq1,each=win)
  indx=indx[1:len]


  ev1sp=split(ev1[1:len],indx)
  ev2sp=split(ev2[1:len],indx)
  tmp=vector('list',length(unique(indx)))
```

```
  for(i in 1:length(tmp)){
    tmp[[i]]=list('strt'=seq1[i],'ev1'=ev1sp[[i]],
                          'ev2'=ev2sp[[i]])
  }

  for(i in 1:length(tmp)){
    tmpccf=ccf(tmp[[i]]$ev1,tmp[[i]]$ev2,plot=F)
    tmpmax=which(tmpccf$acf==max(tmpccf$acf),arr.ind=T)
    tmp[[i]]$ccfmax=tmpccf$acf[tmpmax]
    tmp[[i]]$maxlag=tmpccf$lag[tmpmax]
    rm(tmpccf,tmpmax)
  }

  res=matrix(NA,nrow=length(tmp),ncol=3)
  for(i in 1:length(tmp)){
    res[i,]=c(tmp[[i]]$strt,tmp[[i]]$ccfmax,tmp[[i]]$maxlag)
  }
  dimnames(res)=list(NULL,c('indx','maxcorr','lag'))
  return(res)
}

normPk.r=function(x){
  #function to normalize by max amplitude
  #2011/11/29
```

```
  #x is an element of spec (an event)
  #divide amplitudes by max amplitude

  if(!is.null(x$spec)){
    mmax=max(x$spec,na.rm=T)
    x$normPk=x$spec/mmax
  }elsex$normPk=NULL
  x
}

#####################################################
#do analysis
sta='sta1'

sta1ev=vector('list',length=length(ev))
for(k in 1:length(ev)){
  sta1ev[[k]]=ev[[k]]$sta1
  class(sta1ev[[k]])='rsac'
}
names(sta1ev)=names(ev)

#remove events without arrival times
foo.r=function(x){
```

```
  #x is an element of sta1ev (an event)
  x$a
}
parr=unlist(lapply(sta1ev,foo.r))

for(i in length(sta1ev):1){
  if(is.na(parr[i])) sta1ev[[i]]=NULL
}

#truncate amplitudes to start at P-arrival time
foo.r=function(x){
  #x is an element of sta1ev (an event)
  st=floor(x$a/dt)
  x$amp=x$amp[st:x$N]
}
tmp=lapply(sta1ev,foo.r)
for(k in 1:length(tmp)){
  sta1ev[[k]]$amp=tmp[[k]]
}
class(sta1ev)='rsac'

foo.r=function(x){
  length(x$amp)
}
```

```
minN=min(unlist(lapply(sta1ev,foo.r)))

###########################################
##########################################
#2012/05/25
#cluster time series straight up
#peak normalize time series
foo.r=function(x){
  #get amplitudes into a matrix
  x$amp[1:minN]/max(x$amp[1:minN])
}
sta1evamp=matrix(unlist(lapply(sta1ev,foo.r)),nrow=minN,
             ncol=length(sta1ev))
sta1evamp=t(sta1evamp)  #rows are events, columns are time
tm=seq(0,(minN-1))
dimnames(sta1evamp)=list(names(sta1ev),tm)
tmclust=hclust(dist(sta1evamp)ˆ2,’ward’)

outname=paste(sta,’_tccf_dend’,sep="")
par(mar=c(3,3,1,1),oma=c(4,1,1,1))
plot(tmclust,ylab=’Dissimilarity’,cex=0.7,mgp=c(2,0.5,0),
     tcl=-0.2)
mtext(paste(figpath,outname,".eps     ",sep=""),side=1,
     line=-0.5,outer=T,adj=1,cex=0.6)
```

```
mtext(paste(codepath,"Eg1tccfV1_sta1.r     ",sep=""),side=1,
  line=0.5,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
rm(outname)

#initialize time cluster object (only do once)
#tmcl=vector('list',length=length(ev$ev1))
#names(tmcl)=names(ev$ev1)

nclust=2
tmcl$sta1=cutree(tmclust,k=nclust)

#plot locations color coded by group
colvec2=rep(NA,length(tmcl))
colvec2[tmcl$sta1==1]='blue'
colvec2[tmcl$sta1==2]='green'
colvec2[tmcl$sta1==3]='red'
colvec2[tmcl$sta1==4]='orange'
colvec2[tmp2$sta1==5]='magenta'
match2=match(names(tmcl$sta1),c18locs$ev)

leglab=paste("Group",1:nclust)
```

```
legcol=unique(colvec2)

outname=paste(sta,'_tccf_ENlocs_',nclust,sep="")
par(mfrow=c(1,1),mar=c(3,3,1,1),oma=c(4,1,1,1))
plot(c18locs$east.m[match2],c18locs$north.m[match2],col=colvec2,
    xlab='East (m)', ylab='North (m)', mgp=c(2,0.5,0),tcl=-0.2)
legend(x='topright',legend=leglab, col=legcol, pch=rep(1,3), cex=0.7,
    inset=0.015)

mtext(paste(figpath,outname,".eps     ",sep=""),side=1,line=0,
  outer=T,adj=1,cex=0.6)
mtext(paste(codepath,"Etmfr_ccf_sta2.r     ",sep=""),side=1,
  line=0.75,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
rm(outname)

yrng=c(max(c18locs$depth.m[match2]),min(c18locs$depth.m[match2]))
outname=paste(sta,'_tccf_EZlocs_',nclust,sep="")
par(mar=c(3,3,1,1),oma=c(4,1,1,1))
plot(c18locs$east.m[match2],c18locs$depth.m[match2],
  col=colvec2,xlab='east (m)',ylab='depth (m)',
  mgp=c(2,0.5,0),tcl=-0.2,ylim=yrng)
```

```
mtext(paste(figpath,outname,".eps     ",sep=""),side=1,
  line=0,outer=T,adj=1,cex=0.6)
mtext(paste(codepath,"Etmfr_ccf_sta2.r     ",sep=""),side=1,
  line=0.75,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
rm(outname)

#do cross correlations with group means
#get cluster mean time series
clmeans=vector('list',length=nclust)
names(clmeans)=paste('cl',1:nclust,sep='')
for(i in 1:nclust){
  clmeans[[i]]=apply(sta1evamp[tmcl$sta1==i,],2,mean)
}
#initialize ccf list
tmclccf=vector('list',length=nclust)
names(tmclccf)=paste('cl',1:nclust,sep='')
nlag=30 #corresponds to 0.02 s
tm=(-nlag:nlag)*dt
#populate it
for(i in 1:nclust){
  cat('i=',i,'; ')
```

```
  tmclccf[[i]]=vector('list',nrow(sta1evamp))
  names(tmclccf[[i]])=dimnames(sta1evamp)[[1]]
  for(j in 1:nrow(sta1evamp)){
    cat('j=',j,'; ')
    tmp=ccf(x=clmeans[[i]],y=sta1evamp[j,],lag.max=nlag)
    tmclccf[[i]][[j]]=data.frame(tm,'lag'=tmp$lag,'acf'=tmp$acf)
  }
  rm(tmp)
}
#pull off max ccf and corresponding lag
foo.r=function(x){
  #x is an element of tmclccf[[i]] (an event)
  ind=which(x$acf==max(x$acf),arr.ind=T)
  x[ind,]
}
foo1.r=function(x){
  #x is an element of tmclccf (a cluster mean)
  lapply(x,foo.r)
}
tmp=lapply(tmclccf,foo1.r)
#pull off max
tmccfmax=vector('list',length(tmp))
names(tmccfmax)=names(tmp)
for(i in 1:length(tmp)){
```

```
  tmccfmax[[i]]=cbind(tmcl$sta1,
   matrix(unlist(tmp[[i]]),ncol=3,byrow=T))
  dimnames(tmccfmax[[i]])=list(names(sta1evamp),
    c('cl','tm','lag','maxccf'))
  tmccfmax[[i]]=as.data.frame(tmccfmax[[i]])
}

#plot
outname=paste(sta,'tm_ccfmax',sep='')
mycol=c('blue','green')
colvec=rep(NA,length(tmccfmax$cl1))
ind=tmccfmax$cl1$cl==1
colvec[ind]='blue'
colvec[!ind]='green'
par(mfrow=c(1,2))
plot(tmccfmax$cl1$tm,tmccfmax$cl1$maxccf,col=colvec,cex=0.8,
   xlab='Lag (s)',ylab='Maximum Cross-correlation',
   mgp=c(1.5,0.5,0),tcl=-0.25)
abline(v=0,lty=2,col='gray')
colvec=rep(NA,length(tmccfmax$cl2))
ind=tmccfmax$cl2$cl==1
colvec[ind]='blue'
colvec[!ind]='green'
plot(tmccfmax$cl2$tm,tmccfmax$cl2$maxccf,col=colvec,cex=0.8,
```

```
    xlab='time (s)',ylab='max(ccf)',mgp=c(1.5,0.5,0),tcl=-0.25)
abline(v=0,lty=2,col='gray')
mtext(paste(figpath,outname,".eps     ",sep=""),side=1,line=0,
  outer=T,adj=1,cex=0.6)
mtext(paste(codepath,"Etmfr_ccf_sta2.r     ",sep=""),side=1,
  line=0.75,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
rm(outname)
###########################################
##########################################
#2012/05/25
#cluster spectra straight up

#get spectra for each event, starting with arrival time and series
#length==min(len)
  #if an event does not have an arrival time, spec[event]==NULL

foo.r=function(x){
  #x is an element of sta1ev (an event)
  spectrum(ts(data=x,start=0,frequency=dt),plot=F)
}
sta1spec=apply(sta1evamp,1,foo.r)
```

```
freq=sta1spec[[1]]$freq

#normalize spectra by peak frequency and append to
#each element of spec
tmp=lapply(sta1spec,normPk.r)  #move through events
sta1spec=tmp
rm(tmp)

#cluster using hclust, method='ward'
#1. get spectra into a matrix
foo.r=function(x){
  #x is an element of spec (a station)
  if(!is.null(x$normPk)) x$normPk
}
sta1specmx=matrix(unlist(lapply(sta1spec,foo.r)),ncol=length(freq),
    byrow=T)
dimnames(sta1specmx)=list(names(sta1spec),freq)

#2. cluster
#intialize cluster object (only do this once)
#cl=vector('list',length=length(ev[[1]]))  #number of stations
#names(cl)=names(ev[[1]])  #names are station names

#calculate starting centroids using hclust
```

```
tmp=hclust(dist(sta1specmx)^2,'ward')

#plot dendrogram (station 1 dendrogram plot below)
outname=paste(sta,'fDend_plain',sep="")
par(mar=c(3,3,1,1),oma=c(4,1,1,1))
plot(tmp,cex=0.65,ylab="Dissimilarity (Hz^2)",mgp=c(2,1,0),
  tcl=-0.02,cex.axis=0.8)
mtext(paste(figpath,outname,".eps     ",sep=""),side=1,line=-0.5,
  outer=T, adj=1,cex=0.6)
mtext(paste(codepath,"Etmfr_ccf_sta2.r     ",sep=""),side=1,
  line=0.5,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
rm(outname)

#choose number of clusters
nclust=2
km=cutree(tmp,k=nclust)

#organize output of hclust to be similar to kmeans
match1=match(sta,names(cl))
cl[[match1]]$cluster=km
```

```
colvec2=rep(NA,length(cl$sta1$cluster))
colvec2[cl$sta1$cluster==1]='purple'
colvec2[cl$sta1$cluster==2]='green'
colvec2[cl$sta1$cluster==3]='red'
colvec2[cl$sta1$cluster==4]='orange'
colvec2[cl$sta1$cluster==5]='magenta'
match1=match(names(cl$sta1$cluster),c18locs$ev)

leglab=paste("Group",1:nclust)
legcol=unique(colvec2)

outname=paste(sta,'_fccf_ENlocs_',nclust,sep="")
par(mfrow=c(1,1),mar=c(3,3,1,1),oma=c(4,1,1,1))
plot(c18locs$east.m[match1],c18locs$north.m[match1],col=colvec2,
   xlab='East (m)', ylab='North (m)', mgp=c(2,0.5,0),tcl=-0.2)
legend(x='topright',legend=leglab, col=legcol, pch=rep(1,3),
  cex=0.7,inset=0.015)
mtext(paste(figpath,outname,".eps     ",sep=""),side=1,line=0,
  outer=T,adj=1,cex=0.6)
mtext(paste(codepath,"Etmfr_ccf_sta2.r     ",sep=""),side=1,
  line=0.75,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
```

```
rm(outname)

yrng=c(max(c18locs$depth.m[match1]),min(c18locs$depth.m[match1]))
outname=paste(sta,'_fccf_EZlocs_',nclust,sep="")
par(mar=c(3,3,1,1),oma=c(4,1,2,1))
plot(c18locs$east.m[match1],c18locs$depth.m[match1],col=colvec2,
   xlab='East (m)',ylab='Depth (m)',mgp=c(2,0.5,0),tcl=-0.2,ylim=yrng)
#legend(x='bottomright', legend=c("Group 1","Group 2","Group 3"),
  col=c("blue","green","red"), pch=rep(1,3), cex=0.7, inset=0.015)
abline(h=vel$bdepth[5],lty=2)
mtext(paste(figpath,outname,".eps     ",sep=""),side=1,line=0,
  outer=T,adj=1,cex=0.6)
mtext(paste(codepath,"Etmfr_ccf_sta2.r     ",sep=""),side=1,
  line=0.75,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
rm(outname)

#######################################################
########################################################
########################################################
#plot 2 filtered waves to highlight either shoulder or double peaks
#find events with nearly identical spectra in a low frequency band
```

```
flo=20
fhi=50

outname=paste(sta,'_ev4ev5spec',sep="")
par(mar=c(3,3,1,2),oma=c(4,1,2,1))
plot(sta1specmx[4,1:600],type='l',col='blue',xlab="Frequency (Hz)",
   ylab="Normalized Intensity",mgp=c(2,0.5,0),tcl=-0.2)
lines(sta1specmx[5,1:600],col='green')
abline(v=c(flo,fhi),lty=2)
legend(x='topright',legend=paste('Event',c(4,5)),
  col=c('blue','green'),pch=rep(1,2),cex=0.8,inset=0.01)
mtext(paste(figpath,outname,".eps     ",sep=""),side=1,
  line=0,outer=T,adj=1,cex=0.6)
mtext(paste(codepath,"Etmfr_ccf_sta2.r     ",sep=""),side=1,
  line=0.75,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
rm(outname)

#filter
tmp=vector(mode='list')
tmp[[1]]=sta1ev$ev4
tmp[[2]]=sta1ev$ev5
```

```
class(tmp)='rsac'
x=iir(tmp,fl=flo,fh=fhi,proto="BU",type="BP")
tm=(1:minN)*dt

outname=paste(sta,'_ev4ev5ftrace',sep="")
par(mar=c(3,3,1,2),oma=c(4,1,2,1))
plot(tm,x[[1]]$amp[1:minN]/max(x[[1]]$amp),type='l',
  col='blue',xlab='Time (s)',ylab='Normalized amplitude',
  mgp=c(2,0.5,0),tcl=-0.2)
lines(tm,x[[2]]$amp[1:minN]/max(x[[2]]$amp),col='green')
legend(x='topright',legend=paste('Event',c(4,5)),
  col=c('blue','green'),lty=rep(1,2),cex=0.8,inset=0.01)
mtext(paste(figpath,outname,".eps     ",sep=""),side=1,
  line=0,outer=T,adj=1,cex=0.6)
mtext(paste(codepath,"Etmfr_ccf_sta2.r     ",sep=""),
 side=1,line=0.75,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
rm(outname)

##########################################
#2012/05/24
#calculate mean time series from group 1 correlated with
```

```
#all time series
foo.r=function(x){
  #get amplitudes into a matrix and peak normalize
  x$amp[1:minN]/max(x$amp[1:minN])
}
sta2evamp=matrix(unlist(lapply(sta2ev,foo.r)),nrow=minN,
  ncol=length(sta2ev))
tm=seq(0,(minN-1))
dimnames(sta2evamp)=list(tm,names(sta2ev))
sta2evamp.mn=apply(sta2evamp,1,mean)

foo.r=function(x,y=sta2evamp.mn){
  #x is a row of sta2evamp (an event)
  tmp=ccf(x,y,lag.max=30,plot=F)
}
sta2evtccf=apply(sta2evamp,2,foo.r)
names(sta2evtccf)=names(sta2ev)

foo.r=function(x){
  #x is an element of sta2evtccf (an event)
  length(x$lag)
}
tmp=NULL
for(i in 1:length(sta2evtccf)){
```

```
  tmp=rbind(tmp,sta2evtccf[[i]]$acf)
}
dimnames(tmp)=list(names(sta2ev),seq(-30,30))


outname="sta2_tacfmx"
x=seq(-30,30)*dt
y=1:length(sta2evtccf)
par(mar=c(3,3,1,1),oma=c(4,1,1,1))
image(x,y,t(tmp),col=gray(seq(0,1,length=61)),
  xlab='lag (s)',ylab='',mgp=c(1.5,0.5,0.1),mar=c(4,3,1,1),
    yaxt='n',cex.axis=0.9)
axis(side=2,at=y,labels=names(sta2evtccf),cex.axis=0.7,las=2,
  mgp=c(2,0.5,0),tcl=-0.2)
mtext(paste(figpath,outname,".eps     ",sep=""),side=1,
  line=-0.5,outer=T,adj=1,cex=0.6)
mtext(paste(codepath,"Eg1avgdtvtV2_sta2.r     ",sep=""),side=1,
  line=0.5,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
rm(outname)


tmp2=hclust(dist(tmp),'ward')
tmp3=cutree(tmp2,k=3)
```

```
outname='sta2_tccf_g1dend'
par(mar=c(3,3,1,1),oma=c(4,1,1,1))
plot(tmp2,ylab='dissimilarity',cex=0.7,mgp=c(2,0.5,0),tcl=-0.2)
mtext(paste(figpath,outname,".eps     ",sep=""),side=1,
  line=-0.5,outer=T,adj=1,cex=0.6)
mtext(paste(codepath,"Eg1avgdtvtV2_sta2.r     ",sep=""),side=1,
  line=0.5,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
rm(outname)

clmatch=match(names(tmp3),names(cl$sta2$cluster))
table(tmp3,cl$sta2$cluster[clmatch])

#find events that switch cluster membership
tmp3diff=tmp3-cl$sta2$cluster[clmatch]
table(tmp3diff)
#tmp3diff
ind=tmp3diff!=0
cbind(tmp3[ind],cl$sta2$cluster[ind])
```

```
#original color scheme
#cl$sta2$cluster = 1==1W (blue) 2==2 (green) 3==1E (red)
colvec1=rep(NA,length(cl$sta2$cluster))
colvec1[cl$sta2$cluster==1]='blue'
colvec1[cl$sta2$cluster==2]='green'
colvec1[cl$sta2$cluster==3]='red'
match1=match(names(cl$sta2$cluster),c18locs$ev)

colvec2=rep(NA,length(tmp3))
colvec2[tmp3==1]='blue'
colvec2[tmp3==2]='green'
colvec2[tmp3==3]='red'
match2=match(names(tmp3),c18locs$ev)

outname='sta2_tccf_fccf_locs'
par(mfrow=c(1,2),mar=c(3,3,1,1),oma=c(4,1,1,1))
plot(c18locs$east.m[match1],c18locs$north.m[match1],
  col=colvec1,xlab='east', ylab='north',mgp=c(2,0.5,0),
  tcl=-0.2,cex.axis=0.8)
plot(c18locs$east.m[match2],c18locs$north.m[match2],
  col=colvec2,xlab='east',ylab='north',mgp=c(2,0.5,0),
  tcl=-0.2,cex.axis=0.8)
mtext(paste(figpath,outname,".eps     ",sep=""),side=1,
  line=0,outer=T,adj=1,cex=0.6)
```

```
mtext(paste(codepath,"Eg1tccfV1_sta2.r     ",sep=""),side=1,
  line=0.75,outer=T,adj=1,cex=0.6)
mtext(paste(date(),"    ",sep=""),side=1,line=1.5,outer=T,
  adj=1,cex=0.6)
dev.print(file=paste(figpath,outname,".eps",sep=""))
rm(outname)

q('yes')
```