# PROBABILITY DENSITY FUNCTIONS FOR SECONDARY SEISMIC PHASE ARRIVALS

Eric A. Bergman[1] and E. R. Engdahl[2]

Global Seismological Services[1] and University of Colorado[2]

Sponsored by Defense Threat Reduction Agency

Contract No. DTRA01-02-P-0168

## ABSTRACT

Seismic sources can be easily located with high confidence using only first-arriving P phases when there are large numbers of recording stations, well distributed in azimuth and distance. This is seldom the case, however, for seismic sources of interest to the nuclear explosion monitoring community. In cases of poor azimuthal distribution or small numbers of recording stations, improved accuracy may be gained by incorporating arrival times of secondary phases, but only if phase associations and weighting can be done in a statistically-valid manner. Under this project, we are developing a methodology to use probability density functions (PDFs) for this purpose, with the ultimate goal of making the use of secondary phases a standard part of routine earthquake location procedures. Following Buland (1986) we model the PDFs as a combination of Gaussian and Cauchy distributions, and we explore the usefulness of allowing the relative proportions of these parent distributions to vary for different phases. Sets of PDFs (for secondary phases of interest) are closely tied to the data sets from which they are derived, because the observability of different phases depends on the instrumentation, processing, and analysis procedures employed by contributing stations. A standard methodology for deriving estimates of PDF sets from different data sets is presented, with preliminary results of the analysis of two important arrival time data sets: a subset of ~17,000 well-located seismic sources from the International Seismic Centre (ISC) catalog, processed according to the methodology developed by Engdahl, van der Hilst, and Buland (the EHB standard, Engdahl *et al*, 1998), and an equivalent data set of arrival times from the Center for Monitoring Research. We propose an algorithm for phase association and weighting, based on these PDF sets, and compare locations of test data sets using traditional and PDF-based location methods.

## OBJECTIVE

The objective of this research is to develop methodologies to characterize the distributions of secondary seismic phase arrival times in ways that facilitate their use in earthquake location algorithms. The statistical properties of these distributions should be used both in the association (phase identification) process and in weighting of data for inversion.

## RESEARCH ACCOMPLISHED

### Introduction

Very little research has been accomplished as of the date of submission of this paper, which is only two weeks after the start date of this contract. We have begun assembling the first data set for this study, and this is described below. The remainder of this section provides an introduction and motivation for this research.

This project focuses on several aspects of the earthquake location problem related to the use of secondary phases. These issues are related to the statistical properties of secondary phase arrivals and the optimal handling of those properties in a location program. We are developing methodologies by which these statistical properties can be utilized to give more accurate hypocenters with more accurate uncertainties. Failure to address these problems, while adding secondary phase data to the routine processing pipeline, may well lead to poorer estimates of seismic source parameters than if first-arriving phases alone were used.

Because of the differences in their missions, there are some important differences between the approaches of the earthquake and nuclear explosion monitoring communities in regard to location methodologies. The nuclear explosion monitoring community typically prefers to work with small networks of well-characterized, very high quality seismic stations, especially seismic arrays, while the earthquake community prefers to work with the largest possible set of data, regardless of instrumentation or quality. The nuclear explosion monitoring community is most concerned with location of seismic sources at regional distances, up to about 2000 km, whereas the earthquake community operates at every scale up to global. Earthquake data centers still depend heavily on reported "picks" from cooperating seismic stations, while nuclear explosion monitoring is done with highly trained analysts making the picks of arrival times from waveform data with standardized software and procedures. This additional (and substantial) effort is justified by the need to obtain the highest possible accuracy for small seismic events observed by small numbers of stations.

Despite these differences, the computer algorithms used for this purpose at organizations engaged in monitoring research (Center for Monitoring Research, CMR) or operational monitoring (International Date Centre [IDC], Vienna) are little different from those in use at major earthquake data centers such as the US Geological Survey's National Earthquake Information Center (NEIC), or the International Seismological Centre (ISC) in England. All these programs place more or less severe restrictions on which phases may be utilized other than first-arriving P phases. Other features in common among these programs are a rather crude treatment of phase association, reading errors, and weighting; although these problems exist even for the first-arriving P phase, they are especially severe for secondary phases. In fact, the limited use of secondary phases is, to a large extent, the result of concerns about how to handle these issues.

### Use of Secondary Phases

For a number of years the preferred travel-time tables for nuclear monitoring only listed P waves (Herrin *et al*., 1968). The advantage, of course, is that, being the first arriving phase, it is relatively easy to select an onset time and associate it correctly as the P (or Pn, or PKP, depending on epicentral distance). An advantage of using only P waves is that this method more easily satisfies the requirement of the least squares regression used to solve the location problem, that all data (arrival time picks) are samples of the same random variable. However, even this useful simplification is invalidated when arrival times are selected from different instruments (one of the reasons for the WWSSN).

The use of S waves has now become quite common in local seismic network operations, especially since the ready availability of 3-component instruments. It is feasible because reasonably compatible P and S velocities can be

specified for small regions. Use of S and other secondary phases in location work at regional and global scales has been inhibited by concerns about baseline differences between different phases in the Jeffreys-Bullen (1940) Tables (a particular problem for S waves). This problem is partly resolved by IASPEI91 (Kennett and Engdahl, 1991) and related travel-time models that specifically seek the best *average* one-dimensional (1-D) global travel-time model for all main phases. Still, there are significant baseline offsets observable between different phases along particular source-station paths, reflecting lateral heterogeneity. This problem is worst at regional distances.

Use of secondary phases in the earthquake location algorithm is an obvious and extremely attractive way to increase the amount of data available to locate small seismic events that are observed by small numbers of stations—the most interesting situation in nuclear explosion monitoring. There are several obstacles to be overcome if secondary phases are to be incorporated successfully into routine monitoring work. By "success" we mean that the statistical properties of secondary phase readings are properly carried through the location algorithm. If this is achieved, use of secondary phase data cannot degrade the solution (although it may not improve the solution much). If statistical properties are not properly incorporated, there is a significant chance that the solution will be biased.

First of all, the arrival time of the phase must be picked in a consistent manner. Secondly it must be associated with a phase in order to calculate theoretical travel times. Then, a weight must be assigned. It is tempting to think that the first step is the most important, but in fact there has been little if any research to support that belief. This project deals with the latter two steps. Before we discuss them, however, it is necessary to say something about the characterization of probability density functions for seismic phases.

**Probability Density Functions for Seismic Phases**

It is well known that the probability density functions (PDFs) of seismic phase arrival time residuals can be adequately described by a Gaussian only in the near vicinity of the mean. A full description requires the superposition of a second distribution to account for the “long-tailed-ness” of real seismic residuals. Buland (1986) has shown that P residuals can be well modeled with a combination of a Gaussian and a Cauchy distribution (Evans *et al*., 2000). We will use this approach to describe the PDFs of secondary phases from selected data sets, but we will not assume that the relative proportions of Gaussian and Cauchy are constant among all phases, nor at all distances and depths. In doing this, we must recognize that allowing this variability violates the assumption of least squares concerning a common random variable and must investigate the consequences.

**Phase Association**

Driven by the need to monitor nuclear tests by other countries, earthquake location programs were among the first applications of computers in science, in the early 1960s (Engdahl and Gunst, 1966). These algorithms have changed remarkably little in 40 years. Perhaps the biggest change has been the widespread adoption of the IASPEI91 travel times and associated "tau-p" method of generating travel times (Buland and Chapman, 1983; Kennett and Engdahl, 1991) to replace the Jeffreys-Bullen Tables (Jeffreys and Bullen, 1940). The IASPEI91 model has been superseded by AK135 (Kennett *et al*., 1995) for standard global earthquake location work. The changes in AK135 were adopted to improve agreement between all secondary phases and primary phases. One important consequence of this change is that phase association has been made easier, by making it possible to easily compare theoretical travel times of many candidate phases with an unidentified reading. From a statistical point of view, however, this is far from an adequate solution.

Travel-time "curves", such as those generated using IASPEI91 and the tau-p software, have no reality, being simple representations of complex distributions of reported phase arrival times, with variability arising from mislocation of sources and stations, errors in time-keeping and reporting, uncertainties in picking, instrumentation differences, and—most importantly—lateral variations in Earth's structure. Phase association is a statistical process: regardless of how carefully a pick is made and how well a pick fits a travel-time curve, some doubt must remain. If the statistical nature of phase association is disregarded, the uncertainties of the calculated hypocenter must surely be biased.

The traditional statistic (truncation approach) used for phase association is simply the offset from a baseline, or residual against a travel-time curve. A fuller representation includes a measure of the spread of the distribution, as a function of epicentral distance and source depth, and a measure of the number of hits in the sample from which the

statistics are estimated, also as a function of distance and depth. It is extremely important to choose carefully the dataset from which the statistics are estimated, to reduce the effect of uncontrolled factors. For example, if one wished to determine statistical properties of secondary phase readings for use by the IDC in Vienna, it would be best to work with the dataset of readings in the Reviewed Event Bulletin (REB), and perhaps only a recent subset of the REB if there have been significant changes in procedures or data sources.

As first discussed in Engdahl *et al*. (1998) there is a serious negative consequence of the traditional method of phase association, in which the arrival is assigned to the most likely phase, i.e., the one whose travel-time curve is closest. It is common that two phases will be close to each other, such that their PDFs overlap significantly over a certain distance range. Then the standard procedure is equivalent to truncating each distribution at the point where their probabilities are equal, leaving a longer tail on the other side. The distributions are strongly distance-dependent, becoming more asymmetric as the phases converge. Then, if one were to use these data to estimate a revised "best" (mean) travel time of the phases, the estimate would be biased for both phases, such that they would diverge.

Engdahl *et al*. (1998) developed a solution to this problem for a limited application -- a suite of depth phases (pP, sP, pwP, and PcP) that frequently occur in close proximity. Having estimated the PDFs of these phases, including their relative amplitudes ("observability"), they devised an algorithm to "roll the dice" and assign an arrival to a phase based on the PDFs (Figure 1). Thus, an arrival might be assigned to a phase whose probability is not the highest at that point. Although the location of the earthquake that generated these arrivals will be similarly biased, the greatest concern may be the threat to the statistical integrity of the larger dataset. Because progress in seismology builds on previous work, it is extremely important to avoid introducing systematic biases into the database as a whole.

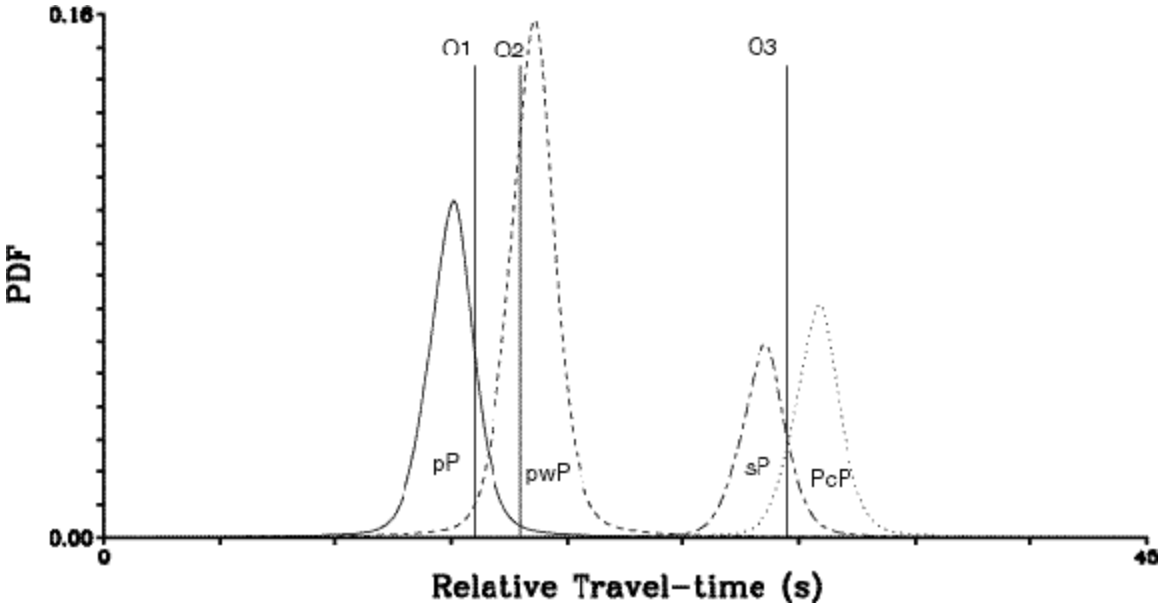


**Figure 1**. Probability density functions (PDFs) for four phases (pP, pwP, sP, and PcP) centered at their theoretical relative travel times from a hypothetical deep event and hypothetical observed phases (O1, O2, and O3) with unknown identifications (from Engdahl *et al*. 1998).

Like Engdahl *et al*. (1998), we will design the association algorithm to simultaneously associate all readings from a given station, forbidding more than one reading to be associated with any given phase. It is also possible for an arrival to be unassociated in this algorithm. In a statistical sense, this is required, as it is entirely possible for a reading to be made from a pulse of energy on a seismogram that has no relationship to the earthquake of interest.

**Weighting of Secondary Phases**

The uncertainty of derived hypocentral parameters is completely controlled by the weights assigned to the data. Computationally, the weights can be anything. Given the formulation of the location problem as a minimization of

the sum of squared errors, and the need to normalize the equation, the only sensible interpretation of weights is in terms of the uncertainty of the readings. In some circumstances it can be acceptable to assign a standard weight to all data; the location itself will not be effected by error in the assigned weight, but the size of the confidence ellipse will be scaled inversely to the weight.

In discussing weights (uncertainties) in location programs, it is important to retain the distinction between uncertainties arising from the picking of the reading from a seismogram and the uncertainty in the theoretical travel time used to calculate the residual. In the nuclear monitoring case (unlike the case of bulletin data), where analysts pick the readings, a direct measure of reading error is possible. Even in the case of bulletin data, it is possible to estimate reading error by multiple event location of clusters of events, because the contribution of the travel-time error is largely removed.

Using this information, it is possible to remove the contribution of reading errors from the PDFs of phases and reveal the PDF associated with lateral heterogeneity in the corresponding dataset. Then this PDF can be properly combined with the actual reading error of individual readings, which will constitute a far more rigorous statistical model than is currently used.

Secondary phase residuals at regional distances exhibit quite large scatter, such that if these readings are weighted inversely to variance, they may contribute almost nothing to the solution. It has been suggested that these phases are simply too difficult to pick. Multiple-event relocation studies (Engdahl and Bergman, 2001) reveal, however, that much of the scatter arises from lateral heterogeneity and from misassociation of phases. Many stations at regional distances consistently pick Pn arrivals from a given cluster (thus minimizing the effect of lateral heterogeneity) with a scatter of no more than a few tenths of seconds, comparable to the scatter of teleseismic arrivals (Figure 2). Remaining large residuals are often attributable to misassociation. The large effects of lateral heterogeneity also have important consequences for the phase association problem, in that arrivals may be far from their "correct" association.

Therefore, to use secondary arrivals in a statistically meaningful way at regional distances, it is essential to begin to unravel the effects of lateral heterogeneity. Major research efforts (e.g., Levshin *et al*, 2001) are addressing the problem of modeling lateral heterogeneity in the crust and upper mantle over large regions, but location algorithms need to be modified to incorporate statistical representations of that heterogeneity. To that end, we propose to investigate the signature of lateral heterogeneity in selected datasets of secondary phase data.

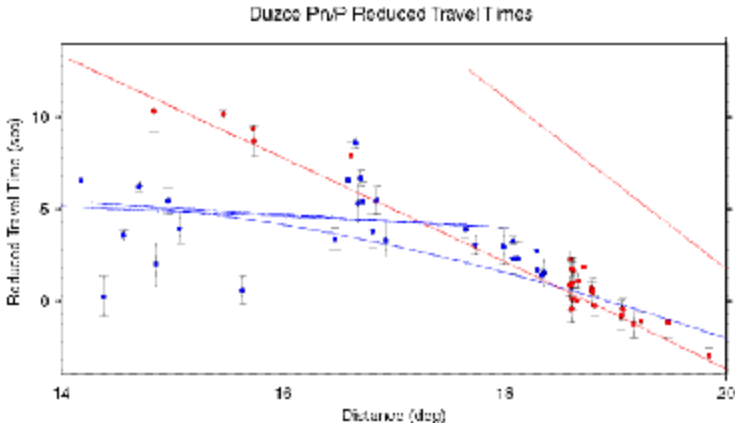


**Figure 2**. Path corrections at regional distance from the 1999 Duzce, Turkey, earthquake, estimated from multiple-event re-location and co-location with known ground truth events (located by a high-density aftershock network). Arrivals were re-associated. Path corrections are plotted against reduced P and Pn travel times for the distance range 14 to 20 degrees. Phases identified in this study as Pn arrivals are plotted in blue and as P arrivals in red. Phase identification in this distance range is complicated by the Pn triplication and by the arrival of a P phase bottoming in the transition zone. In particular, near 18 degrees (keeping in mind the +1 sec bias) phases identified as first arriving Pn could just as easily be associated with phase branches corresponding to P or the Pn cusp. From Engdahl and Bergman (2001).

**A Global Dataset of Secondary Phase Arrivals**

Our first dataset for secondary phase arrivals is derived from the EHB catalog of earthquakes (Engdahl *et al*, 1998). We chose not to use the entire EHB catalog, for three reasons. First, our primary goal of developing new methodologies for analysis of secondary phases would be hindered by the sheer size of the dataset. Second, the quality of the locations of the EHB events is quite variable, leading to excessive scatter in residuals. Finally, the natural distribution of seismic sources in the EHB catalog is quite unbalanced; interaction between the concentration of seismicity and stations in certain regions of the world with the lateral heterogeneity of the Earth will introduce undesirable correlations in the dataset of travel-time residuals.

We addressed all three problems by filtering the catalog with several criteria designed to produce a smaller, but higher quality, dataset of secondary phase arrival times from a set of events that is more evenly distributed around the globe than the raw EHB catalog. A "desirability" index D was calculated for each event, as the product of three terms:

$$\max[(1-(5.6-mag)^2),0]\times\frac{\sqrt{nlat}}{10}\times\max[\frac{(180-sazgap)}{180},0]$$

where *mag* is magnitude, ***nlat*** is the number of secondary phase readings, and ***sazgap*** is the secondary azimuth gap. With these values, events with magnitude less than 4.7 or greater than 6.5 will always have D=0, as will events with no later phase readings and events with a secondary azimuth gap of more than 180 degrees.

Our strategy for choosing a set of optimally distributed events is to divide the Earth into bins approximately 100 km on a side and 75 km in thickness (10 layers in depth). The events in each bin are ranked according to their desirability index D (with a minimum acceptable value of 0.1) and the five highest ranked events are retained for analysis.

The input dataset was a recent update of the EHB catalog, complete from 1964-1999, and pre-screened to remove events with obvious problems. The resulting dataset contained 315,595 events. The selection algorithm reduced this to 17,609 events sampling 6069 bins. There are 5.5 million phase readings associated with these events, an average of 315 per event, but the majority of these are first-arriving P phases.

We have further reduced the size (to 8663 events) of this dataset for our initial analysis, by selecting events with source depths less than 35 km and keeping only readings at epicentral distances less than 20° (Figure 3). The case of regional arrivals from shallow sources is, of course, of greatest relevance to the nuclear explosion monitoring community. We will eventually examine readings from other depth and distance ranges to build a complete picture of the statistical properties of secondary phases.

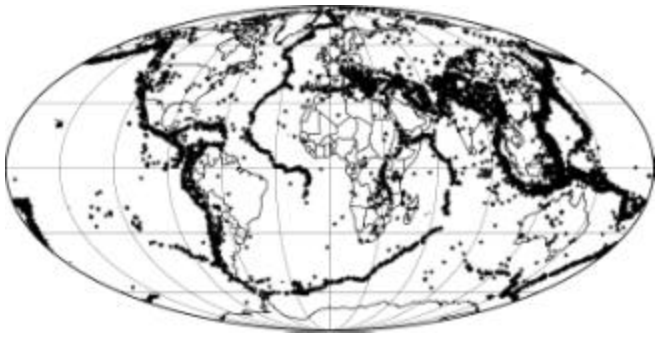

**Figure 3.** More than eight thousand (8,663) events selected for analysis of secondary phases at regional distances from shallow sources. Events shown have depths of less than 35 km and have readings at less than 20° epicentral distance. Most major seismic zones are represented, but a few portions of the mid-ocean ridge system are lost by these criteria.

The nearly-400,000 phase readings in this dataset are shown in Figure 4. The main features of the plot are the combined Pn and Sn arrivals. This kind of display tends to hide the complexities of back branches and triplications, although one can see signs of such features in the P data beyond about 12°.

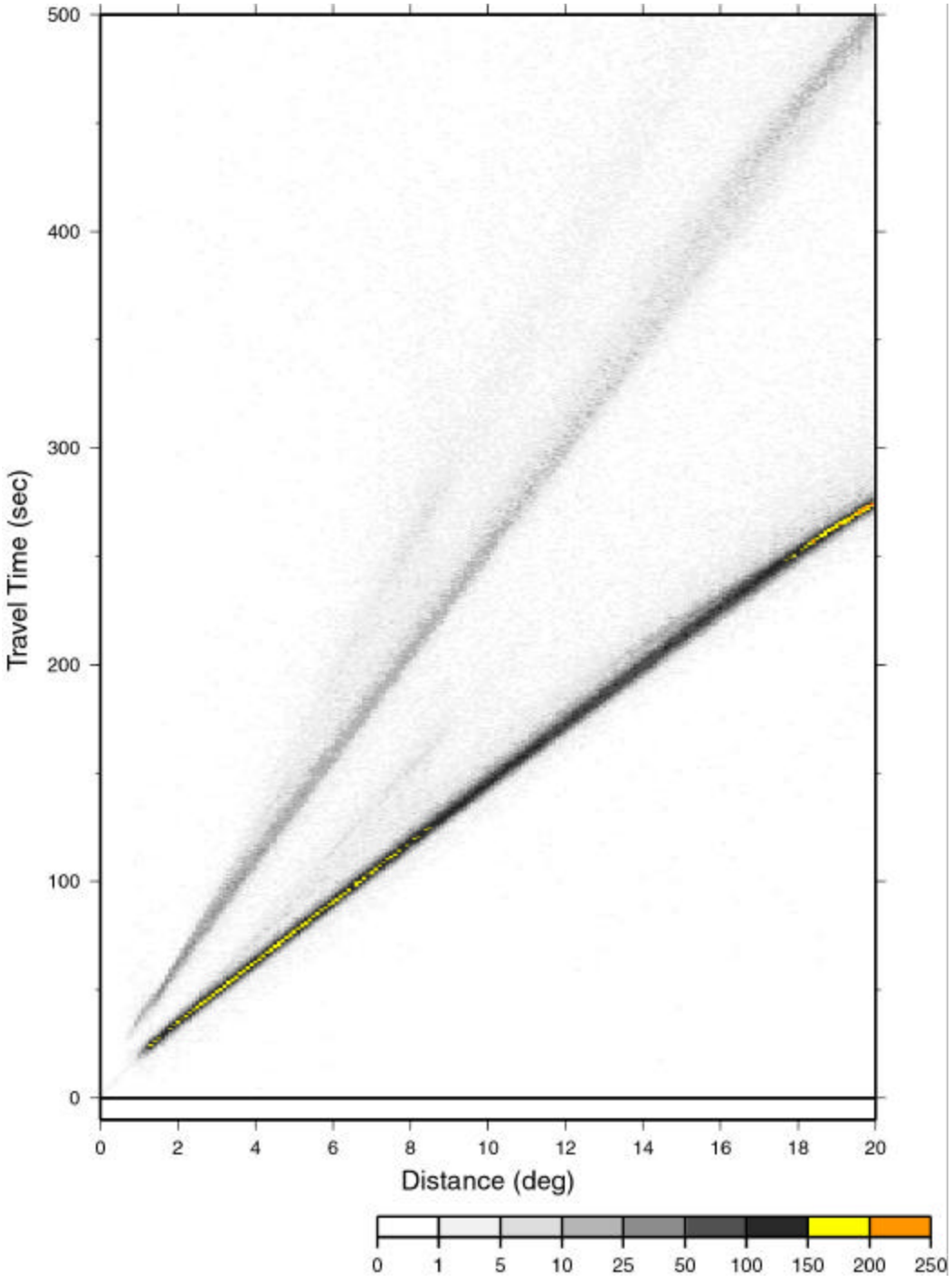


**Figure 4.** Density plot of all phase readings at distances less than 20° for the 8663 shallow (h<35 km) events in Figure 3.

The complexities of these distributions are revealed by density plots over smaller distance ranges, with a reduced time base, such as shown in Figure 5. Note that there appear to be very few observations associated with the back branch of P between 18 and 20°, nor with the cusp of Pn near 18°. However, the back branch of P extending to 14° is relatively well recorded.

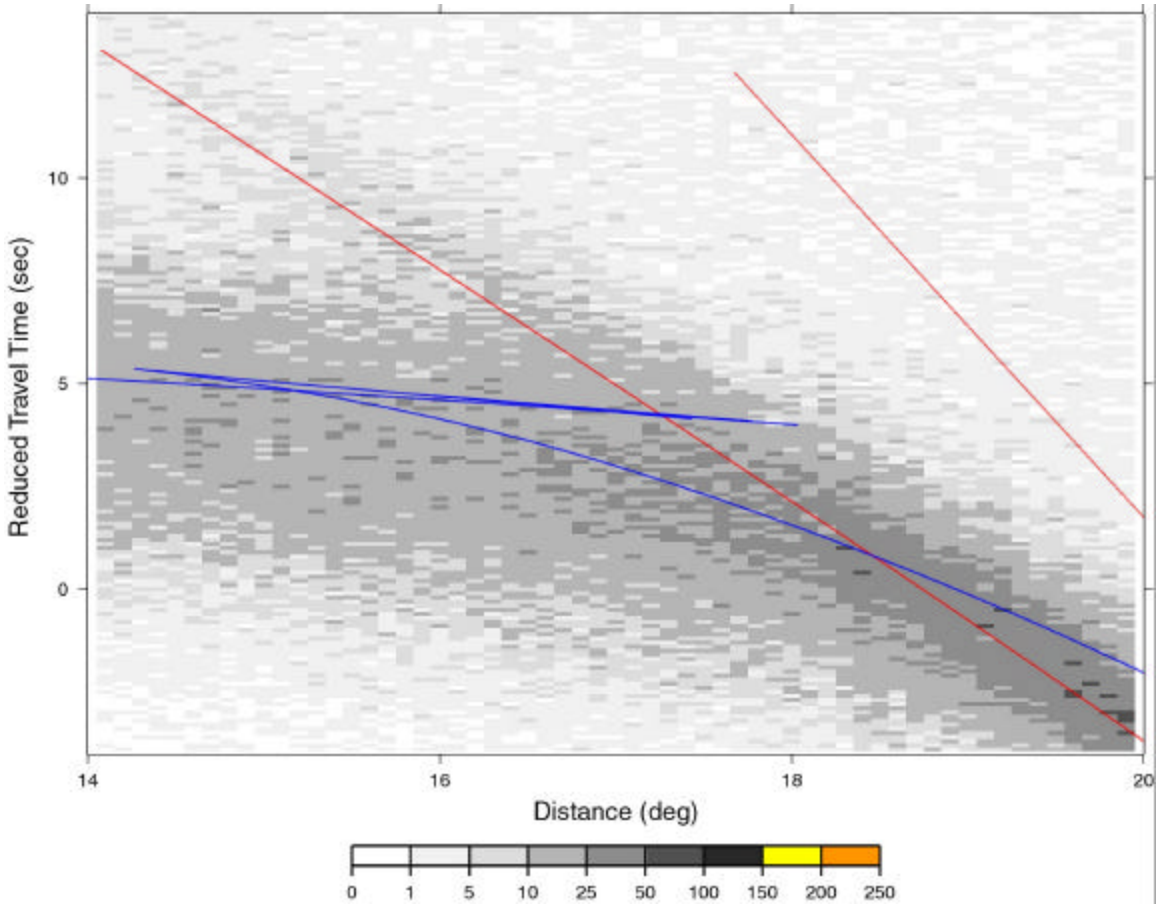


**Figure 5.** Density plot of phase readings near the Pn/P cross-over. Theoretical travel times (AK135) of Pn branches are shown in blue. Theoretical travel times of P branches are shown in red; travel times reduced with a velocity of 8 km/s.

Our analysis of such data will attempt to define PDFs for each significant branch in the data, including the regions where branches cross each other. Other investigations (Figure 2) indicate that much of the scatter observed in Figure 5 is caused by lateral heterogeneity (model error), not the inherent errors in reading these phases. Therefore, the PDFs determined will be strongly influenced by the particular dataset being analyzed. A subset of station-event pairs that sample a common path should display considerably tighter distributions than Figure 5. This will be tested. We should expect that different sets of PDFs will be preferred for different purposes (e.g., global monitoring vs. focused regional monitoring), which highlights the need for standard methodologies for determining PDFs from whatever dataset is of interest.

## CONCLUSIONS AND RECOMMENDATIONS

No conclusions are reached at this preliminary stage of the investigation, except that it would be much more satisfying to be doing the research than writing about the results we hope to achieve.

**REFERENCES**

Buland, R. (1986). Uniform reduction error analysis, *Bull. Seism. Soc. Am.* **76**, 217-230.

Buland, R. and C. Chapman (1983). The computation of seismic travel times, *Bull. Seism. Soc. Am.* **73**, 1271-1302.

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

Engdahl, E. R., R.D. Van der Hilst and R.P. Buland, Global teleseismic earthquake relocation with improved travel times and procedures for depth determination, *Bull. Seism. Soc. Amer*., **88**, 722-743, 1998.

Engdahl, E. R. and E. A. Bergman, Validation and Generation of Reference by Cluster Analysis, Location Workshop (oral presentation) and 23rd Seismic Research Review (poster presentation), Jackson Hole, WY, 1-5 October 2001.

Evans, M., N. Hastings, and B. Peacock (2000), *Statistical Distributions*, 3rd Ed., Wiley & Sons, New York, 221 pp.

Herrin, E., W. Tucker, J. Taggart, D. W. Gordon, and J. L. Lobdell (1968), Estimation of surface focus P travel times, *Bull. Seism. Soc. Am.,* **58**, 1273-1291.

Jeffreys, H. and K. E. Bullen (1940). *Seismological Tables,* British Association for the Advancement of Science, London.

Kennett, B. L. N. and E. R. Engdahl (1991). Travel times for global earthquake location and phase identification, *Geophys. J. Int.* **105**, 429-465.

Kennett, B. L. N., E. R. Engdahl, and R. Buland (1995). Constraints on seismic velocities in the Earth from travel times, *Geophys. J. Int.* **122**, 108-124.

Levshin, A. L., M. H. Ritzwoller, N. M. Shapiro, E. R. Engdahl, M. P. Barmin, and E. A. Bergman (2001), The use of a 3-D model to improve regional event locations (abstract*), Eos Trans. AGU,* **82***(47),* Fall Meet. Suppl., Abstract S12A-0582, 2001.