A REPORT
ON

Classification of Seismic waves into Earthquake and blast waves using Support Vector Machines

BY

ID no.:	Name:	Discipline:
2015A7PS079P	Anshul Jain	B.E Computer Science
2015A7PS028G	Vedang Waradpande	B.E Computer Science

Prepared in partial fulfilment of the
Practice School-I Course
BITS F221/F231/F41

Institute of Seismological Research
(Government of Gujarat, India)

AT

Institute of Seismological Research (ISR), DST
Gandhinagar, Gujrat

A Practice School-I station of

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCES, PILANI
July, 2017

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCES
PILANI (RAJASTHAN)

Practice School Division

Station: Institute of Seismological Research, DST Centre: Gandhinagar, Gujarat
Duration: 53days Date of start: 22/05/17
Date of Submission: 13/07/17
Title of the Project: Classification of Seismic waves into Earthquake and blast waves using Support Vector Machines

ID no. : 2015A7PS028G
Name: Vedang Anand Waradpande
Discipline:B.E. Computer Science & Engineering

Name and designation of the expert: Dr. Santosh Kumar, Scientist D
Name of the PS faculty: Mr Ashish Chittora

Key words:Classification,Earthquake, Support Vector Machines, Machine Learning
Project areas: Machine Learning, Seismology, Geotechnical

Abstract: This project is an effort to create a program which can classify seismic activity into natural Earthquakes and Blast waves. For this purpose, we use Support Vector Machines, a popular Machine Learning algorithm which learns from the data given to it as a training set and predicts if any new instance given to it is an Earthquake or not. We have used two features for this purpose, the logarithm of the Power of event, and the Complexity of the waveform, calculating needed quantities such as the S/P ratio and the Spectral ratio for this purpose.

Signature of Student Signature of PS Faculty
Date: Date:

TABLE OF CONTENTS

ACKNOWLEDGEMENT

First of all, I am very grateful to Institute of Seismological research, Gujarat for giving us an opportunity to work here as interns. Secondly, I would also like to thank Dr. Santosh Kumar our mentor, for their guidance and kind help. I would also like to thank the team in the Datacentre here at Institute of Seismological Research for all their help and support in guiding me upon gathering the required data about the existing system. Lastly, we would like to extend our sincere gratitude towards Mr. Ashish Chittora, our Practice School - I Instructor, for all of her best possible support, guidance and help during my current tenure as intern at the Institute of Seismological Research, Gandhinagar.

INTRODUCTION

The Institute of Seismological Research (ISR) under the Science and Technology Department, Government of Gujarat is functioning from 2006. ISR is the only institute in india fully dedicated to seismological research and is planned to be developed into a premier international institute in few years time.

Earthquake Monitoring Program at ISR

Gujarat seismic network of 60 Broadband Seismographs and 50 Strong Motion Accelerographs is operated since 2006. Data of 36 broadband stations is processed in real-time through VSAT and Auto location software round the clock to determine the epicenter and magnitude of earthquakes within minutes of arrival of the seismic waves and the information is disseminated to administrators for taking appropriate mitigation measures. Space-time pattern of seismicity gives information about newly active faults. One Earthquake Research Center has been started at Bhachau in Kachchh.

Advancing the program

Earthquake detection centres across the state of Gujarat have been installed by the Institute. This is to facilitate the detection of earthquakes so that an appropriate warning can be sent to any city/town nearby. However, the wave signals that are detected don’t get differentiated as those generated by earthquakes or anthropological events such as blasting in mines or other types of explosions. This project is an effort to create a program which can make this distinction with as much accuracy as possible. For this purpose, we use Support Vector Machines, a popular Machine Learning algorithm which learns from the data given to it as a training set and predicts if any new instance given to it is an Earthquake or not. We have used two features, log of the Power of event and Complexity.

PROCEDURE

Selecting the Algorithm

Several methods of varying accuracy have already been developed to achieve the needed goal. However, we choose to go with a machine learning algorithm, since machine learning is a new field which is fast becoming popular within scientific circles, among others. Machine learning algorithms have also been used previously to distinguish between earthquake and blasting waves and have been successful in doing so to varying extents. Also, the accuracy of a machine learning algorithm is known to increase with the amount of data given to it, and hence will only become more and more efficient with time.

There are several popular machine learning algorithms which can be employed to tackle the given problem. Most notably, Artificial Neural Networks (ANNs) have been used for this. There are also other algorithms including Logistic Regression, Linear Regression and Support Vector Machines which can be used.

We have used a support vector machine since they have been known to have a level of accuracy that rivals that of the neural networks and works better with lesser number of features (which we have). It is also known to be better at avoiding overfitting than a neural network.

Datasets

The datasets were in a format called the ASCII format, retrieved from the local database of the Institute. Following is an example of a dataset
:

```
#START_TIME 2017  6  7 18 35  5.000
#SAMP_FREQ   50.00
#NDAT  16751
#STATION_CODE BHI0
#STATION_CHANNEL B  E
      -10052
      -10075
      -10091
      -10103
      -10079
      -10049
      -10068
      -10024
       -9955
       -9923
       -9886
       -9856
       -9836
       -9855
       -9848
       -9839
       -9800
       -9795
       -9863
       -9827
       -9770
       -9739
       -9691
       -9637
       -9573
       -9477
       -9422
       -9363
       -9264
       -9215
       -9187
       -9170
       -9150
       -9165
       -9225
       -9271
       -9293
       -9339
       -9376
       -9370
       -9382
       -9386
       -9388
       -9375
       -9366
       -9404
       -9480
       -9554
       -9575
       -9647
       -9773
       -9856
       -9879
```

Developing the Algorithm

To develop a support vector machine, we need to find features, parameters of the waveform which help differentiate them.

For high energy explosions, the mb:Ms ratio is often used as the dominant feature.

For low energy explosions, a host of features can be used, including:

1. Spectral shape of the whole wave packet
2. P/S ratio
3. Complexity of wave
4. Spectral ratio
5. Ratio of the peaks
6. Power of event (or the logarithm of it)
7. RMS of the time-travel residuals
8. Average amplitude in low and high frequency bands for specific phases
9. Spectral modulation caused by ripple firing waveform inversion
10. Spectral semblance and cross-correlation statistics
11. Measurement of coherency of smoothed spectra at different stations in specific frequency bands.

In our project, we have used two features, viz. Power of event and Complexity, since these two features have been known to give a high accuracy level when used together (Kekovali K. et. al., 2012).

Any event which is not given a label of 'Earthquake' is automatically assumed to be an anthropological event.

A typical SVM needs a kernel to operate, i.e., a function which transposes a low-dimensional space to a high-dimensional space so as to create a much more efficient decision boundary. Examples of kernels are Linear, rbf, Gaussian, etc.

Here, we use a Gaussian kernel, one of the popular kernels, known to produce a boundary with low margins (minimum distance of decision boundary from points bearing either labels) and minimal number of outliers (points of a particular label lying in the area decided to be for another label).

Calculating the Parameters

Complexity:

The complexity is calculated using the ratio of squared integrated spectral amplitudes s(t) of the seismogram from limits (t_1, t_2) to (t_0, t_1), where t_0, t_1 and t_2 were calculated by taking into account the distance of the stations of the stations from the epicentres in the different areas.

Finally, the value of t_0 was taken as the onset time for the P wave, t_1 was taken as 3 secs from t0 and t_2 was taken to be 7 secs from t_0.

Since a definite integral is but the area between the curve and the x-axis, these integrals were calculated as the sum of areas of trapeziums of very small width (say dx).

Log of Power of Event:

The Power of Event is a parameter found by the appropriate combination of the S/P ratio, Complexity and Spectral ratio. This parameter was created especially for this purpose (Kekovali K. et. al. 2012). The Power of event (Pe) is calculated as:

$$Pe = (R_{S/P})^2 \cdot C \cdot (Sr)^2$$

Where $R_{S/P}$ is the S/P ratio, C is the Complexity and Sr is the Spectral ratio of the waveform.

This is directly derived from the fact that plotting the S/P ratio with Sr gives a feature of difference between earthquake and blasting waves, and plotting C with Sr gives another.

The S/P ratio and Spectral Ratio were calculated as follows:

S/P Ratio:

The S/P ratio is the ratio of the amplitudes of the S wave and the P wave, and hence part of the challenge to finding it is calculating the location of these waves.

A program was written for this purpose, which works by comparing the Signal to Noise Ratio (SNR) to a limiting value and then cascading this process through the entire data sample. We know that:

$$SNR = STA \text{ near data-point} / LTA \text{ near data-point},$$

Where STA and LTA are the Short term and Long term averages.

To apply the SNR method, a baseline correction is given to the dataset before use. The actual calculation of the SNR is done for an envelope function instead of the corrected dataset, since adding positive and negative values together would interfere with the calculation of the mean.

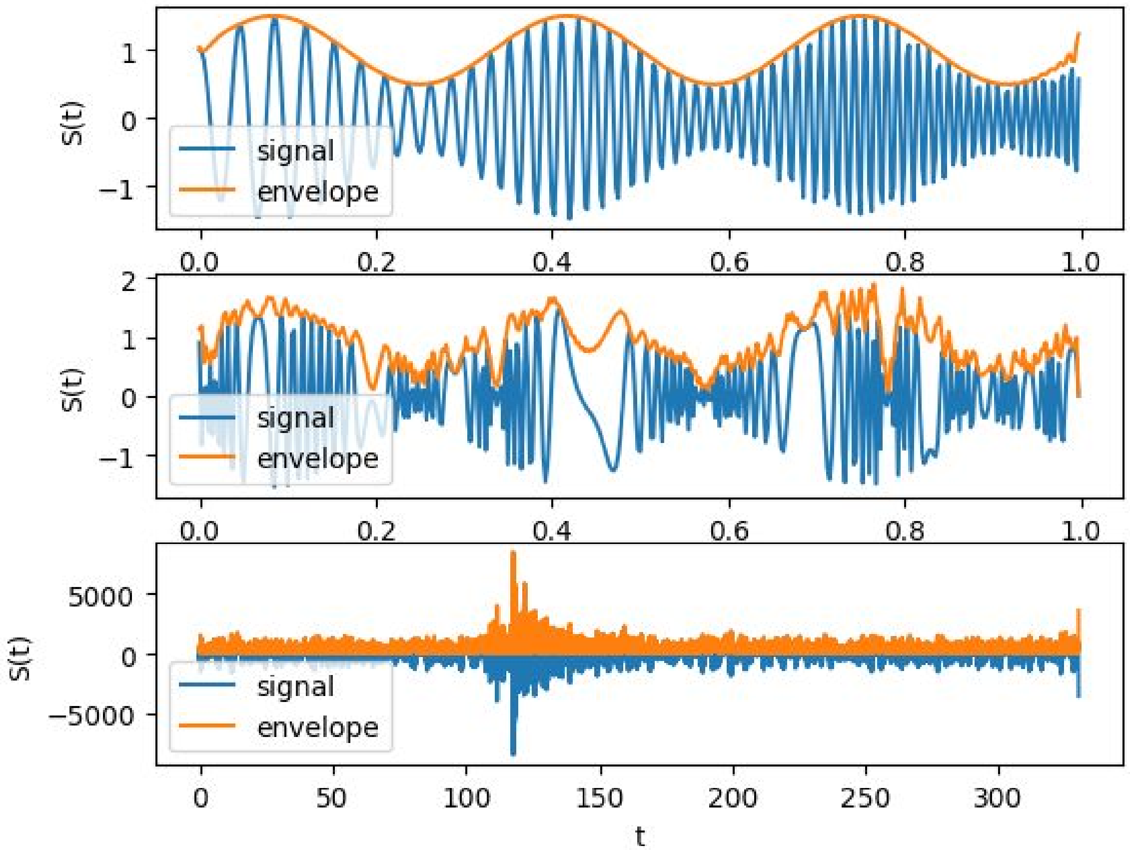
This envelope function is calculated by first finding the hilbert function (Hf(x)) and creating a complex analytic function as follows:

$$A(x) = f(x) + j.Hf(x)$$

An absolute value of this would give the needed envelope function:

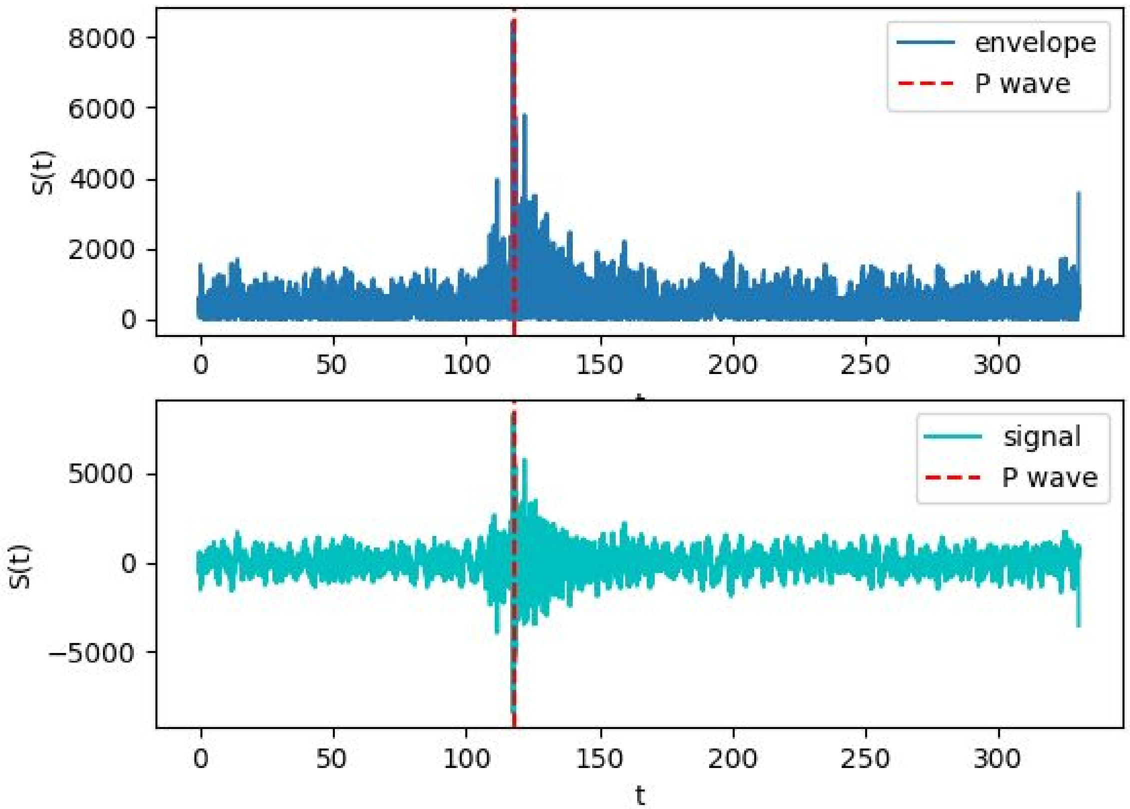
$$E(x) = |A(x)| = \sqrt{f(x)^2 + Hf(x)^2}$$

Some examples of envelope functions (for a periodic function, a non-periodic function, and one of the datasets are shown in the figure below.

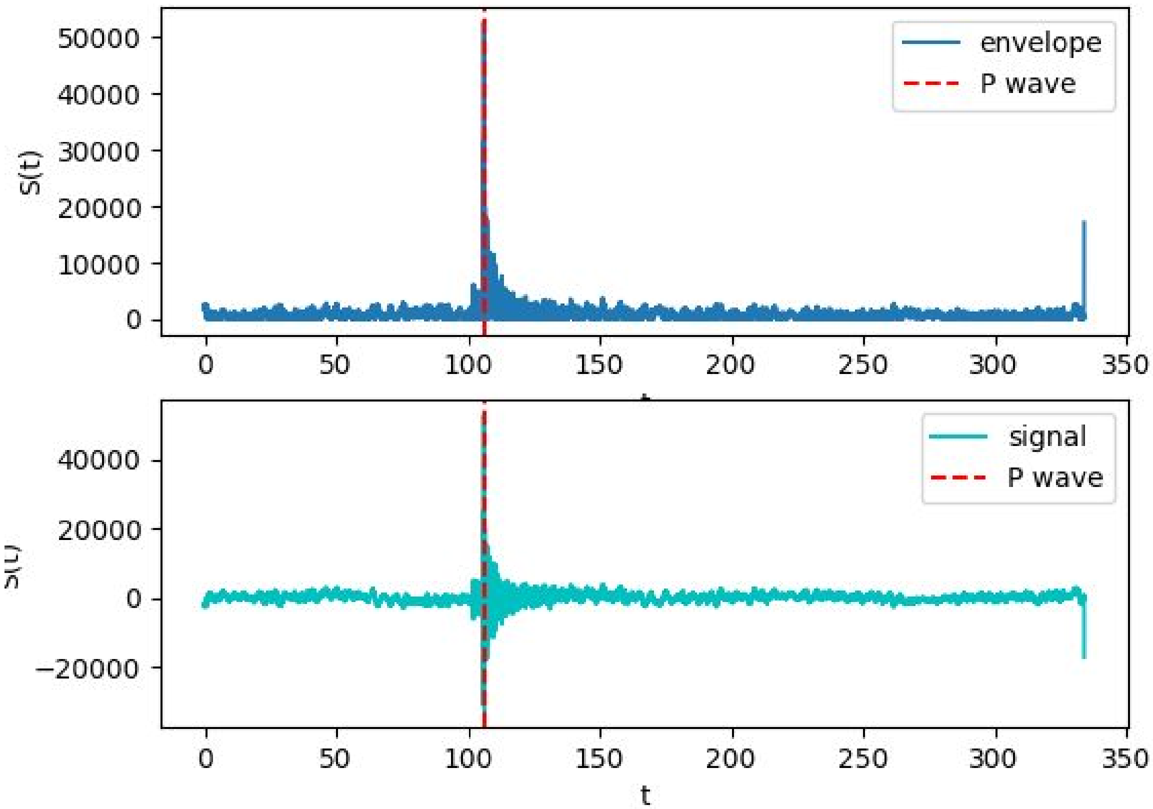


The STA was calculated for 50 values before the current value and the LTA was calculated for 500 values before the current value.

This was seen to give quite good results, as shown in the figures below.



8000
6000
4000
2000
0
S(t)
envelope
P wave
0
50
100
150
200
250
300
t
5000
0
−5000
S(t)
signal
P wave
0
50
100
150
200
250
300
t



Spectral Ratio:

The Spectral ratio (Sr) between the high-frequency (h_1, h_2) and
the low-frequency bands (l_1, l_2) can be represented as Ratio of Integrals (Gitterman and Shapira, 1993):

The low and high frequency bands were calculated by plotting the amplitude vs frequency graph, and finding the ranges of frequency where high energy is observed. This graph is plotted by finding the Fast Fourier Transform (FFT) of the seismogram (Scherbaum and Johnson, 1992).

These integrals were also calculated by summing up the areas of constituent trapeziums of infinitesimal width.

The high and low frequency ranges were decided by taking into account the validity of data according to the Nyquist frequency, 50 Hz in this case, meaning that data from about 0.5 Hz up to 25 Hz can be assumed to be accurate. We have taken (l_1, l_2) as (1, 10) Hz and (h_1, h_2) as (11, 20) Hz, keeping a safe margin to provide more accuracy.

The values of S/P ratio, Complexity and Spectral Ratio were thus substituted to get the corresponding value of Pe. The logarithm (base 10) of this was taken, and the feature taken finally was log Pe.

Structure of Code and Execution of Ideas

The code, written in Python and partly in Shell Script (bash), consists of separate python files for calculating the S/P ratio, Complexity, Spectral ratio and the log of the Power of Event.

A class, Seismogram, was written to make file I/O easier for this specific project. Methods to plot a graph and finding the area were added to this.

There are also bash files which run the program by combining the uses of the different modules and store the output in text files.

Scientific libraries in Python such as Scipy (for a variety of purposes), Scikit-Learn (creating the SVM classifier), Numpy (for a variety of purposes, including finding definite integrals), Matplotlib (plotting graphs), Peakutils (baseline correction) and Pandas (taking input from file) were used where needed.

For full list of programs and their function, see Appendix 1.

RESULTS

We have achieved an accuracy of around 75% using a small dataset where we used 80% of the data for training and 20% for prediction.
We used a Confusion Matrix to calculate the results

Confusion Matrix

	Predicted +	Predicted -
True +	A	B
True -	C	D

Accuracy A+D/A+B+C+D

Confusion Matrix 2

	Predicted +	Predicted -
True +	10	2
True -	3	7

Accuracy A+D/A+B+C+D

Accuracy is 17/22=77.2727%

CONCLUSION

The project has been a great opportunity for acquisition of newer set of skills for us. We have gained new knowledge, developed a new skill-set, and met many new people.

We have shown Machine Learning algorithm can thus be used to differentiate between earthquake and blasting waves, a result which is consistent with past attempts to try to do so. A Support Vector Machine is one such ML algorithm which can be employed.

Plotting the graph of log Pe vs Complexity for a seismic waveform can be very useful in creating a decision boundary (hyperplane) for this algorithm, as this is differentiates the points which denote earthquakes to the points which denote blasts.

Some other useful parameters for classification of earthquake waves and blasting waves can be plotting Complexity vs Spectral Ratio or S/P ratio vs Spectral ratio. There can be other features which can also be used (rms of traveltime residuals, spectral shape of wave packet, etc.).

This project is created assuming we have no knowledge of nearby blast mines or sources of similar anthropological events. Also, the time of the day or night is not taken into consideration. These factors, however can play an important role in this classification. As it is, these can be used in conjunction with this algorithm to achieve a higher accuracy.

Also, the code is designed to work on seismograms of ‘snapshots’ of events, in a text format, which cannot be derived immediately when an event occurs. It should thus be improved upon so that it can work on continuously incoming data. This improved algorithm can thus be used with pre-existing earthquake contingency systems, to immediately classify if an event is an earthquake or a blast, so that needed actions can be taken.

APPENDIX 1

List of Programs And Their Uses

Main python programs:

- Seismogram.py : This is a custom made module containing the class definition of an object of type 'Seismogram' which can be used to read a particular file in a needed way, plot its graph and find the area.
- classifier.py : This is the script code to implementing an SVM.
- complexity.py : This is a module to find the complexity of a particular seismogram.
- pe.py : Program to calculate the log of Power of Event from complexity, S/P ratio and spectral ratio.
- rsp.py : This is a python module to calculate the P/S ratio of the given data by STA / LTA method.
- sr.py : A module to calculate the spectral ratio of the dataset.

Bash programs:

- run.sh : Program to find the feature vector for the classifier and train it.

Relevant text files:

- store.txt : The feature vector is stored in this file.

Resource files:

- 0res_area.py : Resource file for calculating the area in different ways.
- 0res_hilbert.py : Resource file with some examples of implementing the hilbert function available in the module scipy.signal.
- 0res_stationcode.py : Resource file to return the station codes in a text file.

- 0res_stations.txt : Text file to store the station codes of the data samples.
- 0res_readheader.py : Resource python file to fetch and store the headers, displaying the date and time of the event in a text file.
- 0res_callheader.sh : Bash file to fetch and store the headers, displaying the date and time of the event in a text file.
- 0res_headers.txt : File to store the headers.

REFERENCES

1. https://www2.usgs.gov/faq/categories/9839/3338

2. http://calteches.library.caltech.edu/2555/1/berry.pdf

3. http://www.academia.edu/19955422/A_Study_to_Discriminate_Earthquake_from_Mining_Blast_Using_Neural_Network

4. http://www.indiana.edu/~pepp/curriculum/2002/kelley_2002.pdf

5. http://en.rmgc.ro/Content/uploads/uploads_eia_en/anexes/Anexa_NE_Cap4.3_01/Anexa_NE_Cap4.3_01-Assessment%20Study%20on%20the%20Seismic%20Effect%20_En.pdf

6. https://www.ngi.no/eng/Services/Technical-expertise-A-Z/Vibrations-and-earthquakes

7. http://www.sciencedirect.com/science/article/pii/S0098300415300832

8. http://www.iitbhu.ac.in/faculty/min/rajesh-rai/NMEICT-Slope/Pdf/08%20Earthquake%20Analysis.pdf

GLOSSARY

Arrival time: The time at which a particular wave phase arrives at a detector.

Body wave: A seismic wave that can travel through the interior of the earth. P-waves and S-waves are body waves.

Dispersion (wave): The spreading out of a wave train due to each wavelength travelling with its own velocity.

Earthquake: The sudden release of stored elastic energy caused by the sudden fracture and movement of rocks along a fault. Some of the energy released is in the form of seismic waves, that cause the ground to shake.

Epicentre: The point on the earth's surface directly above the focus (hypocentre) of an earthquake.

Frequency: Number of oscillations per unit time; unit is Hertz (Hz),which equals 1 cycle per second.

P wave: Also called primary, longitudinal, irrotational, push, pressure, dilatational, compressional, or push-pull wave. P waves are the fastest body waves and arrive at stations before the S waves, or secondary waves. Their velocity in the crust varies between 5.0 and 7.0 km/s. The waves carry energy through the Earth as longitudinal waves, moving particles in the same line as the direction of the wave. P waves can travel through all layers of the Earth. P waves are generally felt by humans as a bang or thump.

S wave: Also called shear, secondary, rotational, tangential, equivoluminal, distortional, transverse, or shake wave. These waves carry energy through the Earth in very complex patterns of transverse (crosswise) waves. These waves move more slowly than P waves, but in an earthquake they are usually bigger. S waves cannot travel through the outer core because these waves cannot exist in fluids, such as air, water or molten rock.

Seismic Wave: Seismic waves are vibrations generated by sudden movements of rock. After earthquakes occur, the seismic waves propagate from the hypocentre to the surface of the Earth. The speed at which the waves propagate is a function of the nature and type of rock traversed, but generally varies from 1 to 10 km/s. Some waves have a high enough frequency to be audible; others have a very low frequency corresponding to periods of several seconds or minutes.

Seismogram: Recording of ground motions made by a seismograph. This usually is a plot of velocity of the wave vs time or displacement in the Earth vs time.

Surface Waves: Waves that move over the surface of the Earth. Rayleigh and Love waves are surface waves.