17th World Conference on Earthquake Engineering, 17WCEE
Sendai, Japan - September 13th to 18th 2020

On the optimal parameter settings for estimating probabilistic density function of seismic wave amplitude

M. Oshima(1), H. Takenaka(2)

(1) Senior engineer, Shimizu corporation, m.oshima@shimz.co.jp
(2) Professor, Okayama university, htakenaka@cc.okayama-u.ac.jp

Abstract

In seismology and earthquake engineering, sometimes the probability density functions of the amplitude distribution of seismic waveforms are estimated. To estimate a probabilistic density function, at first a histogram is obtained from the amplitude distribution of the seismic waveforms, and a probability density function that fits the shape of the histogram is found.

To obtain a histogram, at first, the range between the maximum and minimum amplitude of the seismic waveform is divided into a certain number of equally spaced classes (bins). Then data samples of the seismic record are assigned to each bin according to their amplitude. In making a histogram, the number of bins should to be determined in advance, and the shape of the resultant probability density function varies according to the number of bins. If the number of bins is too small or too large, the shape of the obtained histogram will be different from the probability density function that the seismic waveform actually follows, and the probability function cannot be estimated correctly. For this reason, when creating a histogram and estimating the probability density function, it is necessary to set the number of bins appropriately.

In practical application, in order to obtain a probability density function, it is necessary to obtain a Gaussian kernel that minimizes an estimation error by cross-validation and to smooth the data using the Gaussian kernel. However, when it is necessary to estimate the probability density function easily and quickly, such as in the case where a large amount of data should be processed in real time, it may be difficult to perform such strict processing. Therefore, we tested several methods, which can determine the number of bins easily and quickly (e.g., Sturges, 1926)[1], to actual seismic waveforms and confirmed the effectiveness of those methods. Those methods determine the number of bins according to the characteristics of the data samples, such as the number of data sample, standard deviation, quartile range, etc. With those methods, we can objectively find the optimum value for the number of bins used in the creation of a histogram. Based on the results of the test, we adopted a method for the calculation of the probabilistic density function with considering the stability and calculation cost.

In the presentation, we will introduce the results of the methods to determine the number of bins we tested and provide some guidelines for finding the probability density function of the seismic waveform easily and quickly by obtaining the histogram. We also show how the probabilistic density functions of the amplitude distribution of seismic records changes as time ticks away with using the method we adopted for the determination of the number of bins.

Keywords: keywords1, Seismic waveforms, amplitude distribution, probabilistic density function, histogram, bin

1. Introduction

In many spheres, PDF of data are estimated and techniques to get PDF has been keenly developed, such as kernel density estimation. In kernel density function estimation, optimal values are assigned to free parameters through cross-validation so that the estimated kernel density explains the data appropriately. However, in seismology and earthquake engineering, sometimes we should process data in real time or should do with enormous amount of data. In such cases, a simple and low-cost way to get PDF is desirable. Up to present, several ways are proposed to determine nbin at low cost. We tested some of those methods and surveyed the applicability of them for real seismic waveforms. The comparison among the PDFs obtained by those methods are done and the relation between twl and Calculation time is also studied.

2. Analysis

We studied the way to set nbins that is applicable under the situation where we can not spare a long Calculation time. So far, some methods to determine nbin of histograms has been proposed. These methos are roughly divided into two types. One is the type that determined nbin referring to only the number of data. The other is the type that determines nbin taking account for the statistical properties of the data.

A typical method of the former type is Sturges (1929). In the method by Sturges (1926), the nbin is given by Eq. (1), where N is the number of data samples within the time window. As examples of the popular methods that falls into the the latter type are Scott(1979)[2] and Freedman and Diaconis(1981)[3]. The method by Freedman and Diaconis (1981) determines the width of a bin by Eq. (2), where h is the width of a bin and IQR is the interquartile range of the data. Then, nbin is obtained by Eq. (4) with the minimum and maximum values of the amplitude of N samples. In the method by Scott (1979), the width of a bin is given by Eq. (3). Then, nbin is obtained by Eq. (4).

$$\text{nbin}=\log_2 N+1 \tag{1}$$

$$\text{h} = \frac{2\text{IQR}}{\sqrt[3]{N}} \tag{2}$$

$$\text{h} = \frac{3.5\sigma}{\sqrt[3]{N}} \tag{3}$$

$$\text{nbin} = (\text{Amax} - \text{Amin})/\text{h} \tag{4}$$

The applicability of the two types of the three methods were examined with the seismic waveform of the event occurred on April 26th, 2016, recorded at station KMMH09. The original acceleration record was once integrated and converted into a velocity record. We used the NS component of the record and the sampling rate of it is 100Hz. We made a comparison of PDFs obtained by the three methods with changing twl. Furthermore, we studied the relation between twl and Calculation time for the three methods.

3. Results and Discussion

There are several popular methods to determine nbin and Sturges (1926), Scott (1979), and Freedman and Diaconis (1981) are among them. Fig.1, Fig.2, and Fig.3 show nbins obtained by Sturges (1926), Scott (1979), and Freedman and Diaconis (1981), respectively. We tested them for twl 0.5s, 1.0s, 2.0s, and 4.0s. Fig.1 shows that nbin does not change with time because Sturges (1926) set nbin only by referring to the

number of data sample. The increase of nbin becomes relatively small when the number of data sample is large.

Fig.2 shows that the number of nbin increases as the number of data sample grows when we use Scott (1979). We also can see nbin varies with the change in amplitude distribution. In Scott (1979), the nbin is determined inversely proportional to the deviation of amplitudes of data samples inside the time window. Therefore the nbin becomes large around P and S waves' arrivals where the deviation of amplitude is small. The shape of nbin's time series is similar to STA/LTA.

Fig.3 shows that the number of nbin increases as the number of data sample grows when we use Freedman and Diaconis (1981). We can understand the nbin by this method is set inversely proportional to IQR of amplitudes of data samples and nbin becomes large around P wave arrival time where IQR of data amplitude is very small compared to the difference of minimum and maximum amplitude.

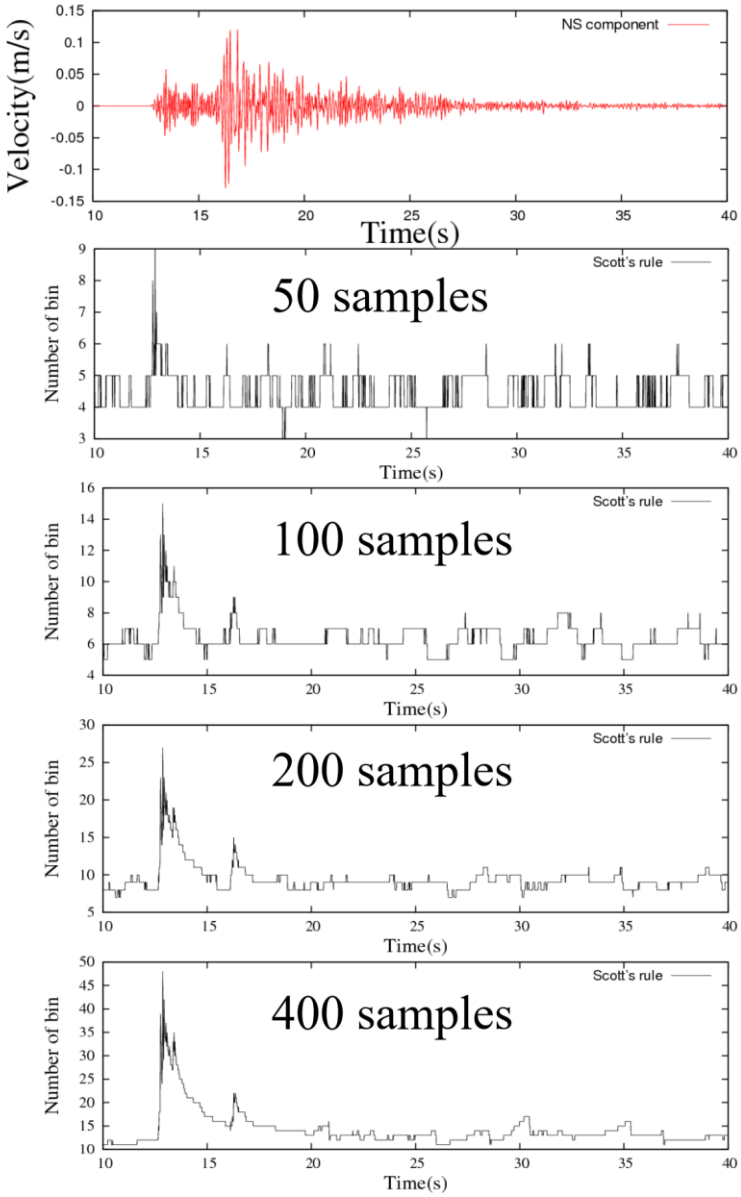

Fig. 1 – The number of bin set by Sturges (1929).

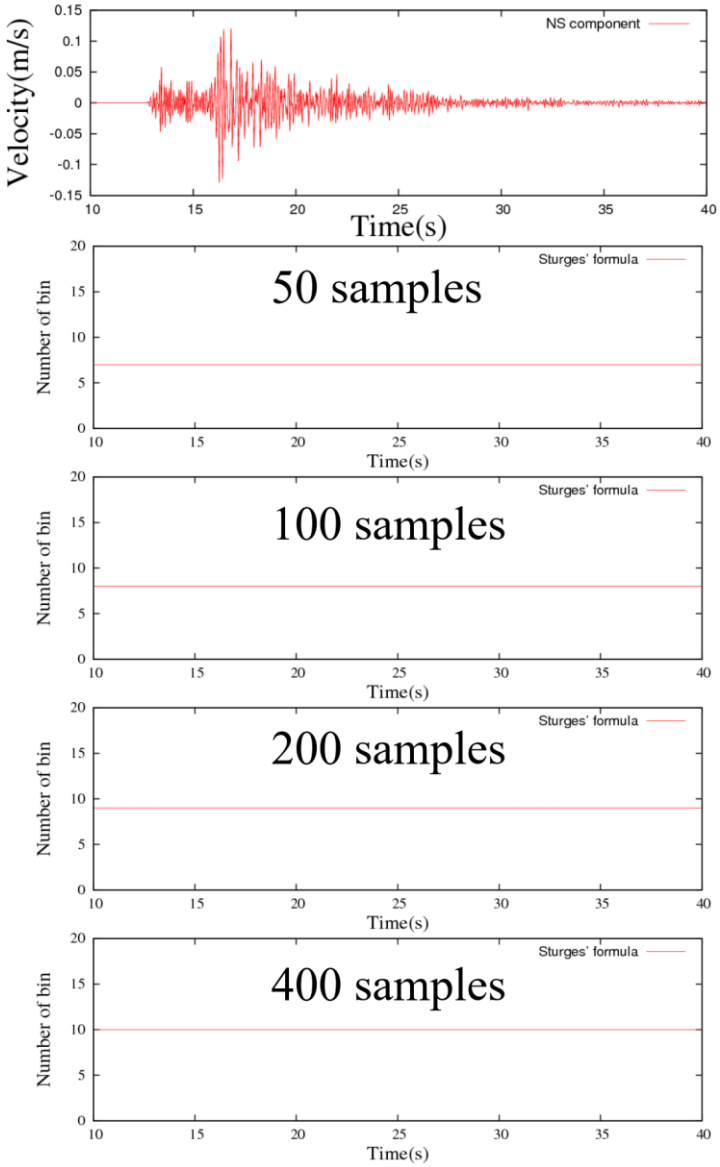

Fig. 2 – The number of bin set by Scott (1979).

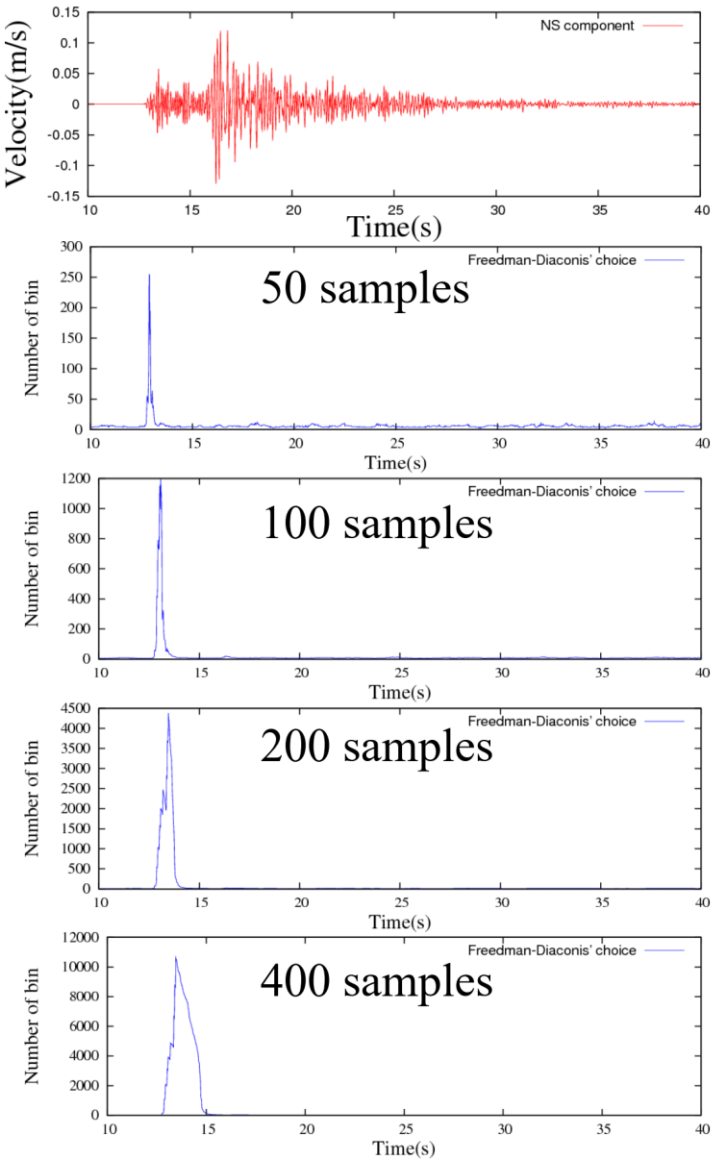

Fig. 3 – The number of bin set by Freedman and Diaconis (1981).

Fig.4 depicts the PDF obtained with setting the nbin by Sturges (1926), Scott (1979)、and Freedman and Diaconis (1981), respectively. The waveform at the top panel is the same as Fig.1-Fig.3. The second to fifth panels from the top are the PDFs for twl equal to 0.5s (50 data samples), 1.0s (100 data samples), 2.0s (200 data samples), and 4.0s (400 data samples), respectively. The leftmost panels are for the time window around P wave arrival time. The second panels from the left are for the time window around S wave arrival time. The rests are for coda wave parts. The horizontal axis is normalized amplitude with a certain shift. In Fig.4, we can see that the shape of the PDF becomes smoother as the number of data sample increase. We also can see the PDFs for P wave, S wave, and Coda waves slightly differ from each other. The cause of the difference maybe stem from the difference among the polarities of those waves.

The nbin by Sturges (1929) tends to be smaller han that by Scott (1979) or Freedman and Diaconis (1981) when the number of data sample is large (n>200). For example, when the number of data sample is 400, the nbin for PDF around P arrival time is 10 by Sturges (1929), where it is 15 by Scott (1979) and 21 by Freedman and Diaconis (1981). Therefore the shape of the PDF has tendency to be rough compared to the

PDF calculated by Scott (1979) or Freedman and Diaconis (1981). On the other hand, the nbin by Sturges (1929) is more than that by the rest two methods when the number of data sample is less or equal to 100.

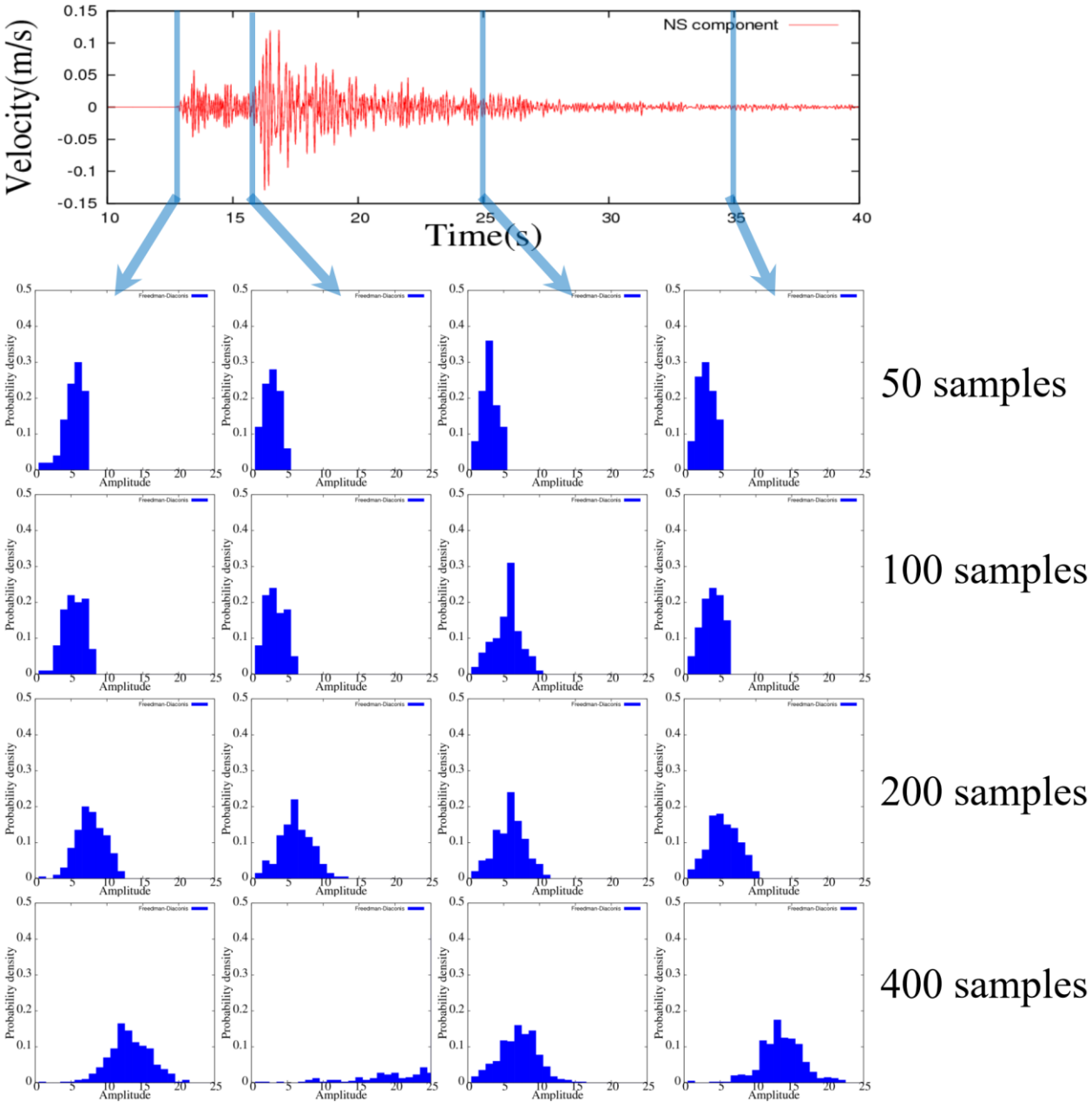

Fig. 4 – The PDF obtained by using Freedman and Sturges (1929).

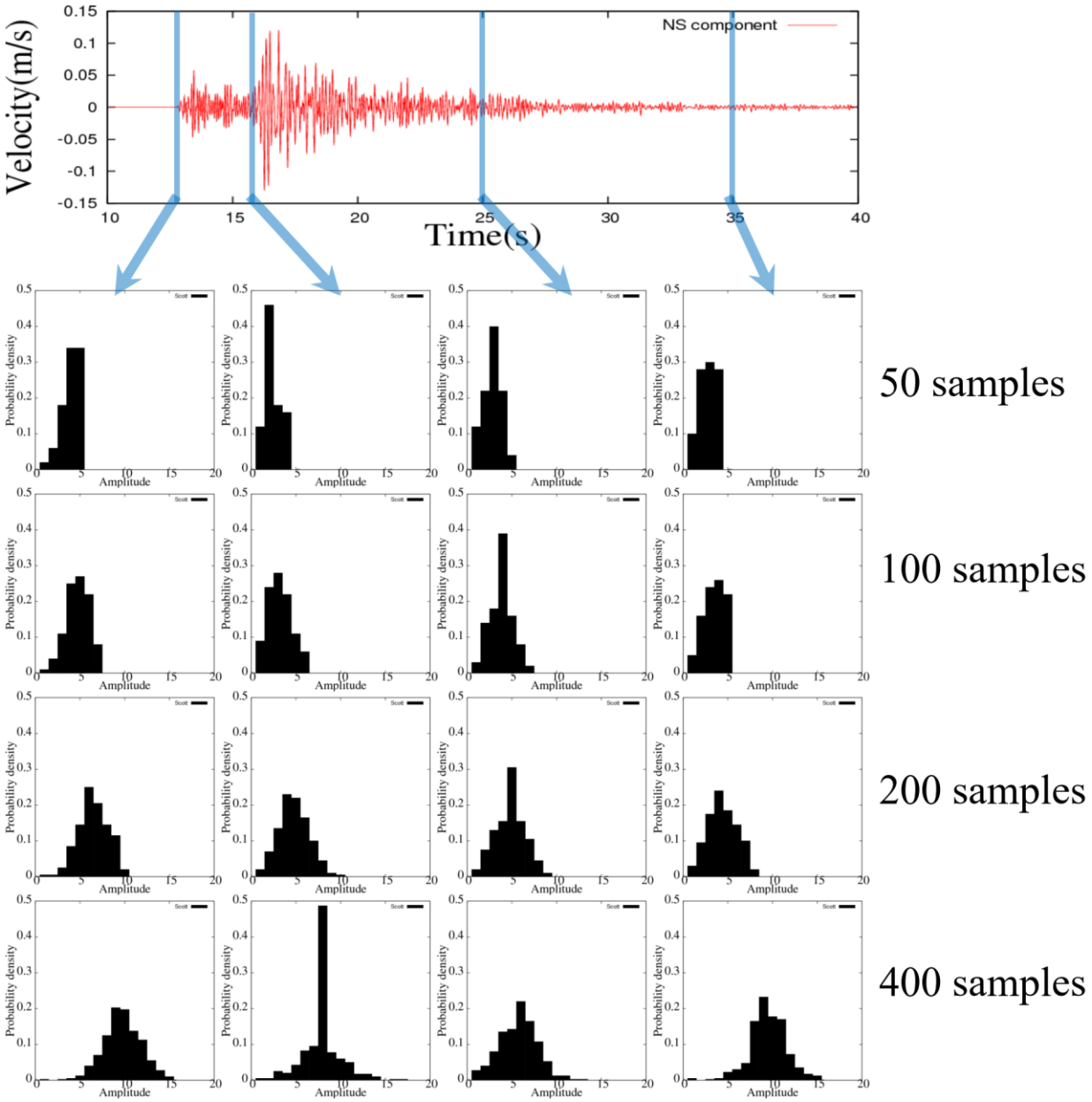

Fig. 5 – The PDF obtained by using Freedman and Scott (1979).

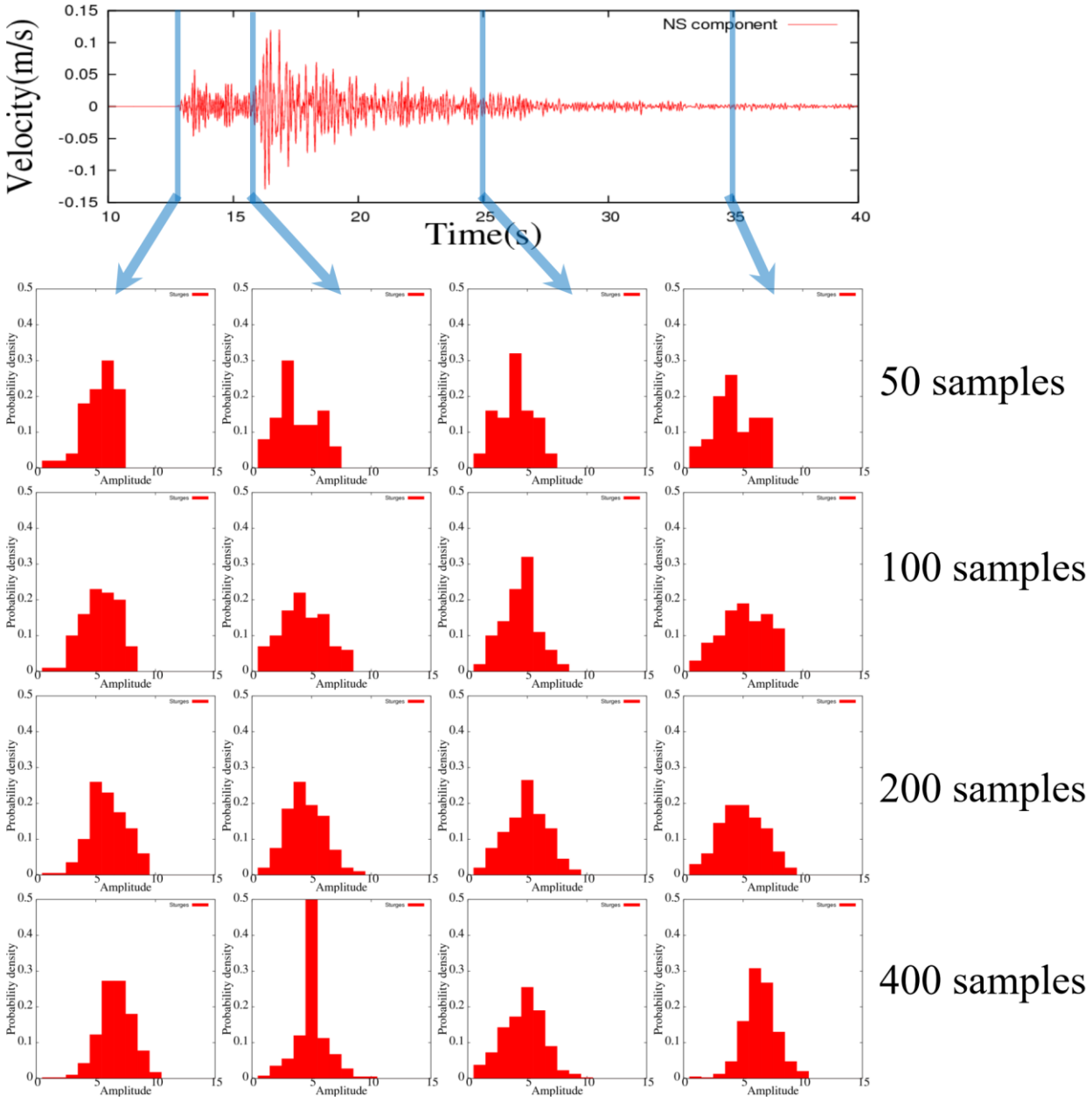

Fig. 6 – The PDF obtained by using Freedman and Diaconis (1981).

The Fig.5 shows that the nbin by Scott (1979) increase as the number of data sample grows and this lead to more detailed shape of PDF. The nbin is as large as Struges (1929) when the number of data sample is less or equal to 200. The nbin by Scottt (1979) is larger than that by Starges (1929) and as same as Freedman and Diaconis (1981) when the number of data sample is 400. The shape of the PDF resembles that obtained using Sturges (1929) in case the number of data is 200 and is more complicated than that by Sturges (1929) in case the number of data sample is 400. This is because the increase of nbin with number of data sample by Sturges (192) becomes dull as the number of data sample becomes large and Scott (1979) set nbin referring to the amplitude deviation.

The Fig.6 shows that the nbin by Freedman and Diaconis(1981) also increase as the number of data sample grows. The nbin by Freedman and Diaconis(1981) is larger than nbins by other two methods in case

the number of data sample is 200 or 400. This is because nbin by Freedman and Diaconis(1981)is set by referring to IQR of data samplitude. The shape of PDF by Freedman and Diaconis(1981) is similar to that by Scott (1979) in case the number of data sample greater than 100 but the spahe is more detailed because nbin is grater. The shape of PDF by Scott (1979) and Freedman and Diaconis(1981) are almost the same in case the number of data is grater than 200.

From Fig. 4, Fig5, and Fig.6, we can understand the applicability of the three method we tested for real seismic waveform in determination of nbin for estimation of PDF. However, we should pay attention to the use of Freedman and Diaconis (1981) because sometimes the nbin changes drastically around P wave arrival time.

In Fig.7, we show the relationship between the twl and the Calculation time for Sturges(1926), Scott(1979)、and Freedman and Diaconis(1981). As the nbin increases, the data samples need to be sorted more finely, and the Calculation time increases. In addition, it takes time to calculate the nbin from the number of data sample in the time window. The calculation time shown in Fig. 7 is the sum of them, and shows the average value of the Calculation time required for 100 times when using Intel (R) Xeon (R) E5-2620 v4, 2.10 GHz. Fig. 7 shows that the Calculation time is longest when Freedman and Diaconis(1981) is used. The Calculation time for Sturges(1929) and Scott(1979) are almost the same and it is proportional to the number of data sample. On the other hand, the Calculation time for Freedman and Diaconis(1981) sharply increases as the number of the data sample grows. This is because the Calculation time necessary to set IQR increases with the increase of the number of data sample. Besides this, more calculation time is needed when Freedman and Diaconis(1981) is used because the nbin by Freedman and Diaconis (1981) is generally more than Sturges (1929) or Scott (1979) and this also boosts Calculation time.

The Calculation time for Freedman and Diaconis(1981) is about three times of that by Sturges(1929) or Scott(1979) when the number of data sample is 300 and that rises up to 6 times when the number of data sample is 400. When real-time processing or huge data needs to be processed, the calculation time becomes a big problem. Therefore, it is necessary to select a method in consideration of the calculation time.

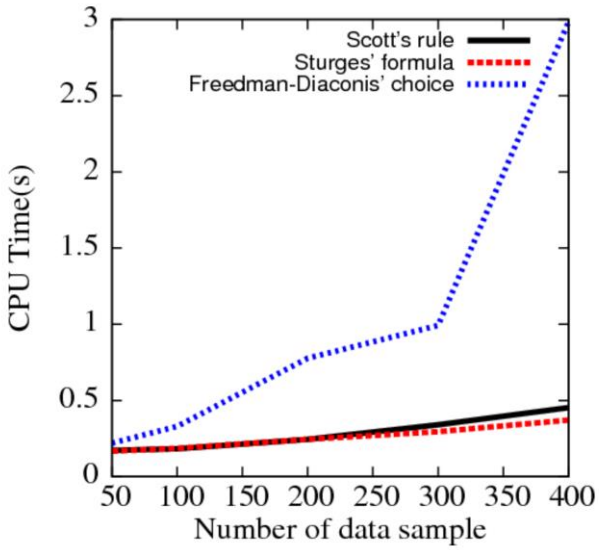

Fig. 7 – The relation between the number of data sample and the calculation time.

4. Conclusions

Assuming the case where real-time processing or processing of huge data is necessary, we studied the setting of the number of bin used in the calculation of the probability density function of seismic wave amplitude. We looked into the method that set the number of bin in making histograms when the number of data sample is already set. The applicability of the methods for real seismic waveforms were investigated. As a result, it was found that the method by Sturges (1926), Scott (1979), and Freedman and Diaconis (1981) was applicable in practical use. However, it should be noted that when Freedman and Diaconis (1981) is used, the number of bin may become a considerably large number near the P wave arrival.

In view of the number of bin obtained by aforementioned three methods, Sturges (1929) gives the largest number when the number of data sample is less than 100. In case the number of data sample is more than 200, then Freedman and Diaconis (1981) gives the largest number for nbin. The shape of the probability density function becomes

As the number of bin rise, the shape of the probability density function is determined in detail, and the shape generally becomes smoother. Since the shape of the probability density function depends on the number of bin, it is necessary to choose an appropriate method depending on the purpose or how detailed the probability density function is to be obtained. Of the three methods tested, it was found that Freedman and Diaconis (1981) gave the largest number of bins and gave a detailed probability density function for more than 200 data samples. It was also found that the shape of the probability density function by Scott (1979) and Freedman and Diaconis (1981) was almost the same when the number of data is more than 200.

The obtained probability density function for P, S, and coda waves are different from each other. This maybe attributed to the difference in the origin and the polarity of those waves. Our study on the relation between the time window length (the number of data sample) and calculation showed that the calculation time by Freedman and Diaconis (1981) is the longest and that by Sturges (1929) and Scott (1979) are almost the same. The calculation time is almost proportional to the number of data sample when Sturges (1929) or Scott(1979) is used. In contrast, the calculation time for Freedman and Diaconis (1981) drastically increases as the number of data sample increase.

As clarified in this study, the shape of the probability density function differs depending on what method is used to set the number of bin, and the calculation time also significantly changes. In application to seismic waveforms, we should select a method according to the purpose while considering how detailed the shape of the probability density function is to be obtained and how much calculation time is acceptable.

5. Acknowledgements

In the present study, we used strong motion data of KiK-net (Aoi et al., 2000)[4]. Generic Mapping Tools (Wessel et al., 2013)[5] was used to create some figures.

6. References

[1] Sturges H A (1926): The choice of a class interval. Journal of the american statistical association, **21** (153), 65-66.

[2] Scott D W (1979): On optimal and data-based histograms. Biometrika, 66, 605-610.

[3] Freedman D, Diaconis P (1981): On the histogram as a density estimator: L 2 theory. Probability theory and related fields, 57, 453-476.

[4] Aoi S, Kunugi T, Fujiwara H (2004): Strong-motion seismograph network operated by NIED: K-NET and KiK-net. Journal of Japan association for earthquake engineering, 4, 65-74.

[5] Wessel P, W H F Smith, R Scharroo, J. F. Luis, F. Wobbe (2013): Generic Mapping Tools: Improved version released, EOS Trans. AGU, 94, 409-410.

