Stochastic Hydrology and Hydraulics 11: 423-431 © Springer-Verlag 1997

Effect of class-interval size on entropy

V.P. Singh
Department of Civil and Environmental Engineering, Louisiana State University, USA

Abstract: The value of Shannon entropy for a given set of data depends on the class interval chosen to compute the relative frequency of each class. For three data sets, expressed in dimensional as well as nondimensional form, the entropy value was computed for different class-interval sizes. Entropy was found to decrease with increasing class interval as well as with increasing sampling interval. It is suggested that these intervals should be selected with care.

Introduction

For a continuous random variable X, the Shannon (1948) entropy H(x) is expressed as

$$H(x) = -\int_a^b f(x) \ln f(x) dx \tag{1}$$

where f(x) is the probability density function (pdf) of X, which varies from a lower limit a ($> -\infty$) to an upper limit b(>0) and whose specific value is denoted by x. For a given set of data H is computed with use of a discrete equivalent of equation (1) which can be expressed as follows: Let X take on values $x_1, x_2, \cdots, x_m$ covering the entire range. We then choose a class interval Δx and arrange the values in different class intervals. Counting the number of values in each class interval and dividing this number by the total number of values, the relative frequency or probability associated with each class interval is obtained. Therefore, equation (1) can be expressed as

$$H(x) = -\sum_{i=1}^{n} f(x_i) \Delta x \ln \left[f(x_i) \frac{\Delta x}{\Delta x} \right] \tag{2}$$

where n is the number of class intervals of size Δx and $f(x_i)$ is the relative frequency associated with the ith class interval. Denoting $f(x_i)\Delta x$ by P_i, equation (2) can be expressed as

$$H(x) = -\sum_{i=1}^{n} P_i \Delta x \ln \left[\frac{P_i}{\Delta x} \right] \tag{3}$$

Equation (3) can be expressed as

$$H(x) = -\sum_{i=1}^{n} P_i \ln P_i + \sum_{i=1}^{n} P_i \ln \Delta x \tag{4}$$

Recalling that

$$\sum_{i=1}^{n} P_i = 1 \tag{5}$$

equation (4) becomes

$$H(x) = -\sum_{i=1}^{n} P_i \ln P_i + \ln \Delta x \tag{6}$$

In usual practice equation (6) is written as

$$H(x) = -\sum_{i=1}^{n} P_i \ln P_i \tag{7}$$

implying that $\ln \Delta x$ is equal to zero. Strictly speaking, this is true only if $\Delta x = 1$. If $\Delta x < 1$, $\ln \Delta x$ will be negative and will reduce the value of H. On the other hand, for $\Delta x > 1$, the value of H will increase. The question then arises as to the contribution of $\ln \Delta x$ to the value of H(x). To that end we write equation (6) as sum of two parts

$$H(x) = H_1 + H_2, \; H_1 = -\sum_{i=1}^{n} P_i \ln P_i, \; H_2 = \ln \Delta x \tag{8}$$

For different class interval sizes it may be interesting to examine the following ratios

$$R_1 = \frac{H_1}{H}, \; R_2 = \frac{H_2}{H}, \; R_3 = \frac{H_2}{H_1} \tag{9}$$

On the other hand, if we start with the discrete representation of entropy given by equation (7) as was done by Shannon (1948) then its continuous counterpart in discrete form takes the form:

$$H(x) = -\sum_{i=1}^{n} f(x_i) \Delta x \ln [f(x_i) \Delta x] \tag{10}$$

where $P_i = f(x_i)\Delta x$. Equation (10) can be expressed as

$$H(x) = -\sum_{i=1}^{n} f(x_i) \Delta x \ln f(x_i) - \ln \Delta x \tag{11}$$

For Δx being small and replaced by dx, equation (11) becomes

$$H(x) = -\int f(x) \ln f(x) dx - \ln \Delta x \tag{12}$$

The right side of equation (12) consists of two parts: the first part is the same as equation (1) and the second part is $-\ln \Delta x$. Thus,

$$H(x) = H_3 + H_4, \; H_3 = -\int f(x) \ln f(x) dx, \; H_4 = -\ln \Delta x \tag{13}$$

Note that $H_4 = -H_2$, and $H_3 = H_1$. In terms of ratios,

$$R_4 = \frac{H_3}{H} = \frac{H_1}{H}, \; R_5 = \frac{H_4}{H} = -\frac{H_2}{H}, \; R_6 = \frac{H_4}{H_3} = -\frac{H_2}{H_1} \tag{14}$$

It should again be emphasized that in usual practice the term H_4 is neglected, implying that it is zero and that Δx is unity. This of course is not true. These and related issues are discussed in what follows.

2 Application

Three sets of discharge data (in cfs units) were used. The measured data available were at unequal time intervals. First, the measured data were transformed logarithmically. Then, the data were interpolated to equal time intervals. The interpolation was linear, based on two observed values-one previous and the other subsequent to the reference value. Seven different sampling time intervals were chosen: 1200 s (20 min), 1800 s (30 min), 3600 s (1h), 5400 s (1.5h), 7200 s (2h), 10,800 s (3h), and 18,000 s (5h). With use of these time intervals, the effect of sampling or observational time was evaluated. Clearly, there would be more data points in a data set if the sampling time was small. Corresponding to each time interval there would be one data subset. The original data set corresponding to the smallest time interval was thus transformed to seven data subsets.

A data subset corresponding to a given time interval was arranged in different classes for a chosen class-interval size. The probabilities associated with the class intervals were computed. Then the value of entropy was computed for the sample (data subset) corresponding to the specified class-interval size. Both H_1 and H_2 as well as the ratios R_1, R_2, and R_3 were computed. Similarly H_3 and H_4 and ratios R_4, R_5, and R_6 were also computed.

The above analysis is based on dimensional (units) values of the data. To eliminate the effect of units the data were nondimensionalized by dividing each value by the mean of the sample data. Then the computations were performed as in the dimensional case for all three data sets. It was found that the value of entropy in the dimensionless case was the same as in the dimensional case. This was also true with the pattern of variation with sampling interval as well as with class interval. Therefore, the results of calculations will be presented for dimensional data only.

Comparing equations (6) and (12) it is seen that depending upon the class interval size the two equations may yield very different values of entropy as shown in Figures 1 and 2 and Table 1 for data set 1. This is because the term, ln Δx, may assume a significant value. In general, this is seen from columns 5 and 6 as well as columns 8 and 11. Indeed this term may become dominant in the extreme when Δx is large. In general, equation (12) is used most frequently in practice.

The value of entropy decreased with increasing class interval. This was true for all sampling intervals of set 1, as shown in Figure 1 and Table 1, if the entropy was defined by equation (12). This was also true if the entropy was defined by equation (6), except for the 5-h subset, where the entropy value increased with increased class interval. Comparing the decrease of entropy of different subsets, it is observed that the rate of decrease with class interval depended on the subset and was not the same

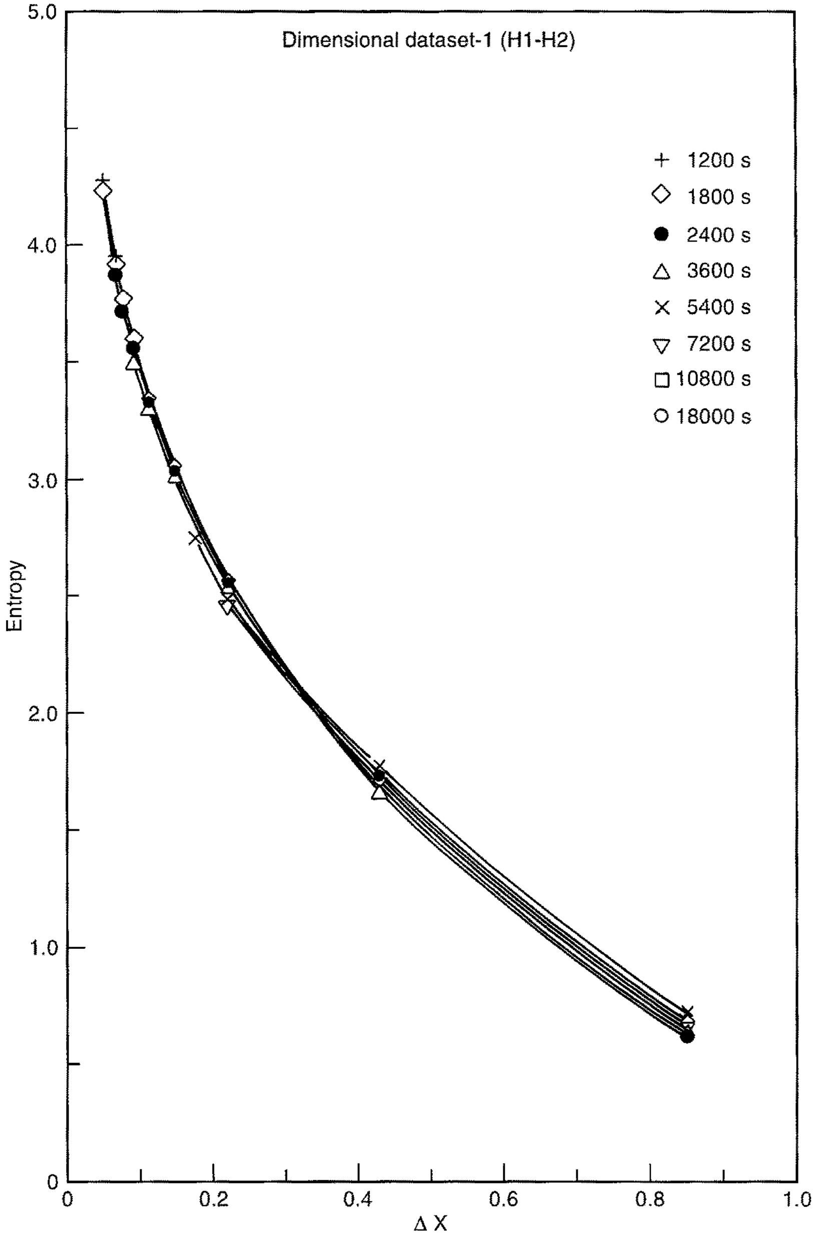

Figure 1. Variation of entropy defined by equation (12) with class interval size for data set

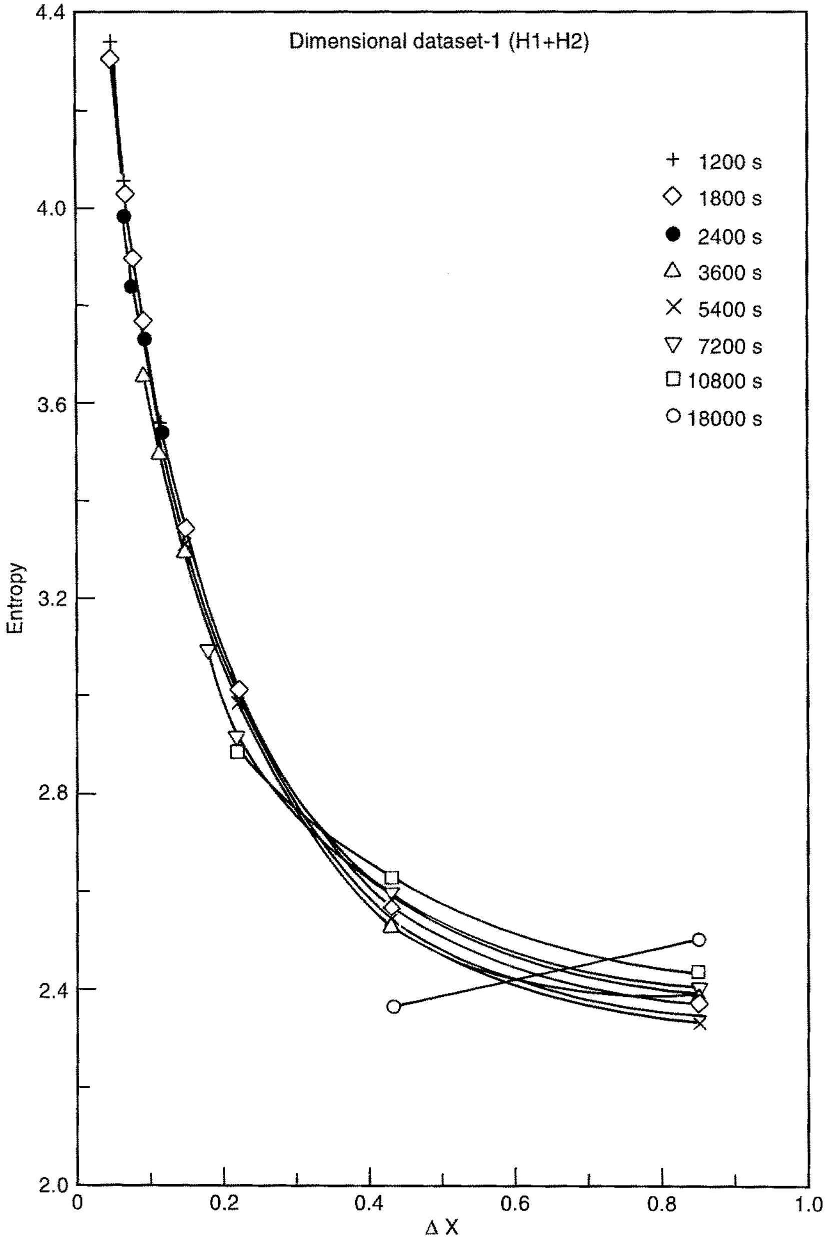

Figure 2. Variation of entropy defined by equation (6) with class interval size for data set 1.

Table 1. Variation of entropy with class-interval size for sampling interval = 1200 secs for data set 1 (dimensional)

No. of delta X	delta X	H1	H2	Entropy H_a=H1+H2	Entropy H_b=H1-H2	R1=H1/H_a	R2=H2/H_a	R3=H2/H1	R4=H1/H_b	R5=H2/H_b	R6=H2/H1
100	0.043	4.301	0.043	4.344	4.258	0.990	0.010	0.010	1.010	-0.010	-0.010
70	0.061	3.999	0.061	4.060	3.937	0.985	0.015	0.015	1.016	-0.016	-0.015
60	0.072	3.826	0.072	3.898	3.755	0.982	0.018	0.019	1.019	-0.019	-0.019
50	0.086	3.680	0.086	3.766	3.595	0.977	0.023	0.023	1.024	-0.024	-0.023
40	0.107	3.454	0.107	3.561	3.347	0.970	0.030	0.031	1.032	-0.032	-0.031
30	0.143	3.198	0.143	4.341	3.055	0.957	0.043	0.045	1.047	-0.047	-0.045
20	0.215	2.779	0.215	2.994	2.565	0.928	0.072	0.077	1.084	-0.084	-0.077
10	0.429	2.112	0.429	2.541	1.683	0.831	0.169	0.203	1.255	-0.255	-0.203
5	0.858	1.488	0.858	2.346	0.630	0.634	0.366	0.577	2.353	-1.363	-0.577

Variation of entropy with class-interval size for sampling interval = 1800 secs for data set 1 (dimensional)

No. of delta X	delta X	H1	H2	Entropy H_a=H1+H2	Entropy H_b=H1-H2	R1=H1/H_a	R2=H2/H_a	R3=H2/H1	R4=H1/H_b	R5=H2/H_b	R6=H2/H1
100	0.043	4.268	0.043	4.311	4.226	0.990	0.010	0.010	1.010	-0.010	-0.010
70	0.061	3.975	0.061	4.035	3.914	0.985	0.015	0.015	1.016	-0.016	-0.015
60	0.071	3.830	0.071	3.901	3.759	0.982	0.018	0.019	1.019	-0.019	-0.019
50	0.085	3.683	0.085	3.768	3.598	0.977	0.023	0.023	1.024	-0.024	-0.023
40	0.107	3.437	0.107	3.544	3.331	0.970	0.030	0.031	1.032	-0.032	-0.031
30	0.142	3.203	0.142	3.345	3.061	0.958	0.042	0.044	1.046	-0.046	-0.044
20	0.213	2.791	0.213	3.004	2.578	0.929	0.071	0.076	1.083	-0.083	-0.076
10	0.426	2.143	0.426	2.570	1.717	0.834	0.166	0.199	1.248	-0.248	-0.199
5	0.853	1.516	0.853	2.369	0.663	0.640	0.360	0.563	2.286	-1.286	-0.563

Variation of entropy with class-interval size for sampling interval = 2400 secs for data set 1 (dimensional)

No. of				Entropy	Entropy						
delta X	delta X	H1	H2	H_a=H1+H2	H_b=H1-H2	R1=H1/H_a	R2=H2/H_a	R3=H2/H1	R4=H1/H_b	R5=H2/H_b	R6=H2/H1
70	0.061	3.928	0.061	3.989	3.867	0.985	0.015	0.015	1.016	-0.016	-0.015
60	0.071	3.771	0.071	3.842	3.700	0.982	0.018	0.019	1.019	-0.019	-0.019
50	0.085	3.651	0.085	3.736	3.565	0.977	0.023	0.023	1.024	-0.024	-0.023
40	0.107	3.430	0.107	3.537	3.324	0.970	0.030	0.031	1.032	-0.032	-0.031
30	0.142	3.173	0.142	3.315	3.031	0.957	0.043	0.045	1.047	-0.047	-0.045
20	0.213	2.776	0.213	2.989	2.563	0.929	0.071	0.077	1.083	-0.083	-0.077
10	0.426	2.105	0.426	2.532	1.679	0.832	0.168	0.202	1.254	-0.254	-0.202
5	0.852	1.481	0.852	2.333	0.628	0.635	0.365	0.576	2.257	-1.357	-0.576

Variation of entropy with class-interval size for sampling interval = 3600 secs for data set 1 (dimensional)

No. of				Entropy	Entropy						
delta X	delta X	H1	H2	H_a=H1+H2	H_b=H1-H2	R1=H1/H_a	R2=H2/H_a	R3=H2/H1	R4=H1/H_b	R5=H2/H_b	R6=H2/H1
50	0.085	3.575	0.085	3.661	3.490	0.977	0.023	0.024	1.024	-0.024	-0.024
40	0.107	3.397	0.107	3.503	3.290	0.970	0.030	0.031	1.032	-0.032	-0.031
30	0.142	3.144	0.142	3.286	3.002	0.957	0.043	0.045	1.047	-0.047	-0.045
20	0.213	2.803	0.213	3.016	2.590	0.929	0.071	0.076	1.082	-0.082	-0.076
10	0.426	2.101	0.426	2.527	1.675	1.831	0.169	0.203	1.254	-0.254	-0.203
5	0.852	1.535	0.852	2.387	0.682	0.643	0.357	0.555	2.249	-1.249	-0.555

Variation of entropy with class-interval size for sampling interval = 5400 secs for data set 1 (dimensional)

No. of				Entropy	Entropy						
delta X	delta X	H1	H2	H_a=H1+H2	H_b=H1+H2	R1=H1/H_a	R2=H2/H_a	R3=H2/H1	R4=H1/H_b	R5=H2/H_b	R6=H2/H1
30	0.142	3.172	0.142	3.314	3.030	0.957	0.043	0.045	1.047	-0.047	-0.045
20	0.213	2.773	0.213	2.986	2.560	0.929	0.071	0.077	1.083	-0.083	-0.077
10	0.426	2.168	0.426	2.594	1.742	0.836	0.164	0.197	1.245	-0.245	-0.197
5	0.852	1.542	0.852	2.395	0.690	0.644	0.356	0.552	2.234	-1.234	-0.552

Variation of entropy with class-interval size for sampling interval = 7200 secs for data set 1 (dimensional)

No. of				Entropy	Entropy						
delta X	delta X	H1	H2	H_a=H1+H2	H_b=H1-H2	R1=H1/H_a	R2=H2/H_a	R3=H2/H1	R4=H1/H_b	R5=H2/H_b	R6=H2/H1
25	0.170	2.925	0.170	3.095	2.754	0.945	0.055	0.058	1.062	-0.062	-0.058
20	0.213	2.706	0.213	2.919	2.493	0.927	0.073	0.079	1.085	-0.085	-0.079
10	0.426	2.168	0.426	2.594	1.742	0.836	0.164	0.197	1.245	-0.245	-0.197
5	0.852	1.555	0.852	2.407	0.703	0.646	0.354	0.548	2.213	-1.213	-0.548

Variation of entropy with class-interval size for sampling interval = 10800 secs for data set 1 (dimensional)

No. of				Entropy	Entropy						
delta X	delta X	H1	H2	H_a=H1+H2	H_b=H1-H2	R1=H1/H_a	R2=H2/H_a	R3=H2/H1	R4=H1/H_b	R5=H2/H_b	R6=H2/H1
20	0.213	2.678	0.213	2.891	2.465	0.926	0.074	0.080	1.086	-0.086	-0.080
10	0.426	2.204	0.426	2.630	1.778	0.838	0.162	0.193	1.239	-0.239	-0.193
5	0.851	1.581	0.851	2.433	0.730	0.650	0.350	0.538	2.166	-1.166	-0.538

Variation of entropy with class-interval size for sampling interval = 7200 secs for data set 1 (dimensional)

No. of				Entropy	Entropy						
delta X	delta X	H1	H2	H_a=H1+H2	H_b=H1-H2	R1=H1/H_a	R2=H2/H_a	R3=H2/H1	R4=H1/H_b	R5=H2/H_b	R6=H2/H1
10	0.426	1.941	0.426	2.367	1.515	0.820	0.180	0.219	1.281	-0.281	-0.219
5	0.851	1.649	0.851	2.500	0.798	0.660	0.340	0.516	2.067	-1.067	-0.516

for all subsets. There appeared to be considerable loss of information when the sampling interval was increased from say 1200 s to 3600 s. For a subset the loss of information was greatly accentuated when the class interval was quadrupled. This was seen to be true for all three data sets.

For the second data set, the entropy values followed a similar pattern as in case of the first data set, except that entropy began to increase' for more sampling intervals when the class interval reached a particular value. In case of data set 3, the increase in entropy beyond a certain class interval was more pronounced. The point of increase, however, depended upon the sampling interval. For lack of space figures and tables for these two data sets are not included.

3 Conclusions

The following conclusions are drawn from this study: (1) The value of entropy decreases as the class interval increases. The rate of decrease, however, is not uniform. (2) In general the entropy value decreases as sampling interval increases meaning loss of information. The rate of decrease, however, is not uniform. (3) The class interval size has a pronounced effect on entropy. Therefore, its accurate selection is important. (4) The sampling interval exercises a great deal of significant influence on entropy and therefore its optimum selection is important for design of sampling schemes. (5) The dimensionality of data has little influence on entropy.

References

Shannon, C.E. 1948: A mathematical theory of communication. Bell System Technical Journal 27, 379-423 and 623-659