# Multiscale Entropy Analysis (MSE)

Madalena Costa, Ary L. Goldberger and C.-K. Peng
Beth Israel Deaconess Medical Center, Boston, USA

A detailed description of the multiscale entropy algorithm and its application can be found in:

- Costa M., Goldberger A.L., Peng C.-K. Multiscale entropy analysis of biological signals. *Phys Rev E* 2005;**71**:021906.
- Costa M., Goldberger A.L., Peng C.-K. Multiscale entropy analysis of physiologic time series. *Phys Rev Lett* 2002; **89**:062102.

Please cite these publications when referencing this material, and also include the standard citation for PhysioNet:

- Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PCh, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**(23):e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215]; 2000 (June 13)

Readers of this tutorial may also wish to read:

- Costa M, Peng C-K, Goldberger AL, Hausdorff JM. Multiscale entropy analysis of human gait dynamics. *Physica A* 2003;**330**:53-60.

The software described in this tutorial is available here.

## 1 Background

Multiscale entropy (MSE) analysis [1, 2] is a new method of measuring the complexity of finite length time series. This computational tool can be applied both to physical and physiologic data sets, and can be used with a variety of measures of entropy. We have developed and applied MSE for the analysis of physiologic time series, for which we prefer to estimate entropy using the sample entropy (SampEn) measure [3]. SampEn is a refinement of the approximate entropy family of statistics introduced by Pincus [4]. Both have been widely used for the analysis of physiologic data sets [5, 6].

Traditional entropy measures quantify only the regularity (predictability) of time series on a single scale. There is no straightforward correspondence, however, between regularity and complexity. Neither completely predictable (e.g., periodic) signals, which have minimum entropy, nor completely unpredictable (e.g., uncorrelated random) signals, which have maximum entropy, are truly complex, since they can be described very compactly. There is no consensus definition of complexity. Intuitively, complexity is associated with "meaningful structural richness" [7] incorporating correlations over multiple spatio-temporal scales.

For example, we and others have observed that traditional single-scale entropy estimates tend to yield lower entropy in time series of physiologic data such as inter-beat (RR) interval series than in surrogate series formed by shuffling the original physiologic data. This happens because the shuffled data are more irregular and less predictable than the original series, which typically contain correlations at many time scales. The process of generating surrogate data destroys the correlations and degrades the information content of the original signal; if one supposes that greater entropy is characteristic of greater complexity, such results are profoundly misleading. The MSE method, in contrast, shows that the original time series are more complex than the surrogate ones, by revealing the dependence of entropy measures on scale [8, 9, 10, 11, 12].

The MSE method incorporates two procedures:

1. A "coarse-graining" process is applied to the time series. For a given time series, multiple coarse-grained time series are constructed by averaging the data points within non-overlapping windows of increasing length, $\tau$ (see Figure 1). Each

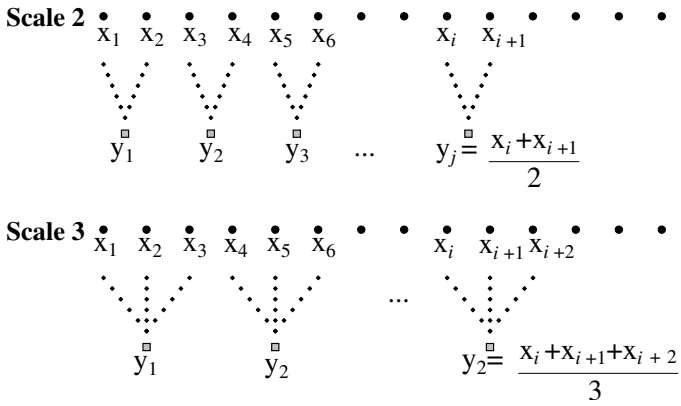


Figure 1: Schematic illustration of the coarse-graining procedure for scale 2 and 3. Adapted from reference [8].

element of the coarse-grained time series, $y_j^{(\tau)}$, is calculated according to the equation:

$$y_j^{(\tau)} = 1/\tau \sum_{i=(j-1)\tau+1}^{j\tau} x_i \tag{1}$$

where $\tau$ represents the scale factor and $1 \leq j \leq N/\tau$. The length of each coarse-grained time series is $N/\tau$. For scale 1, the coarse-grained time series is simply the original time series.

2. SampEn is calculated for each coarse-grained time series, and then plotted as a function of the scale factor. SampEn is a "regularity statistic." It "looks for patterns" in a time series and quantifies its degree of predictability or regularity (see Figure 2).

# 2 MSE analysis of simulated white and 1/f noise

Figure 3 presents the MSE results for simulated uncorrelated (white) and long-range correlated (1/f) noise. Note that for scale one, a higher value of SampEn is obtained for white noise time series than for 1/f time series. Although the value of entropy for the coarse-grained 1/f series remains almost constant for all scales, the value of entropy for the coarse-grained white noise time series monotonically decreases, such that for scales above 4, it becomes smaller than the corresponding values for 1/f noise. In contrast with the conclusions drawn from single-scale entropy-based analyses, the MSE results are consistent with the fact that, unlike white noise, 1/f noise contains correlations across multiple time scales and is, therefore, more complex than white noise [13].

# 3 Software for MSE analysis

Download `mse.c`, the C language source for a program that performs multiscale entropy analysis. The program can be compiled using any ANSI/ISO C compiler, and should be linked to the C math library (it uses only the `sqrt` function from that library). For example, using the freely available GNU C compiler, `mse.c` can be compiled into an executable `mse` by the command:

```
gcc -o mse -O mse.c -lm
```

## Preparing data for MSE analysis

In this tutorial, we illustrate the use of `mse` to analyze time series of intervals between consecutive heart beats (RR intervals). RR interval lists as used by `mse` are in text format, consisting of one column (the RR intervals). Interval lists in this format can be prepared from beat annotation files using ann2rr. Use a command of the form:

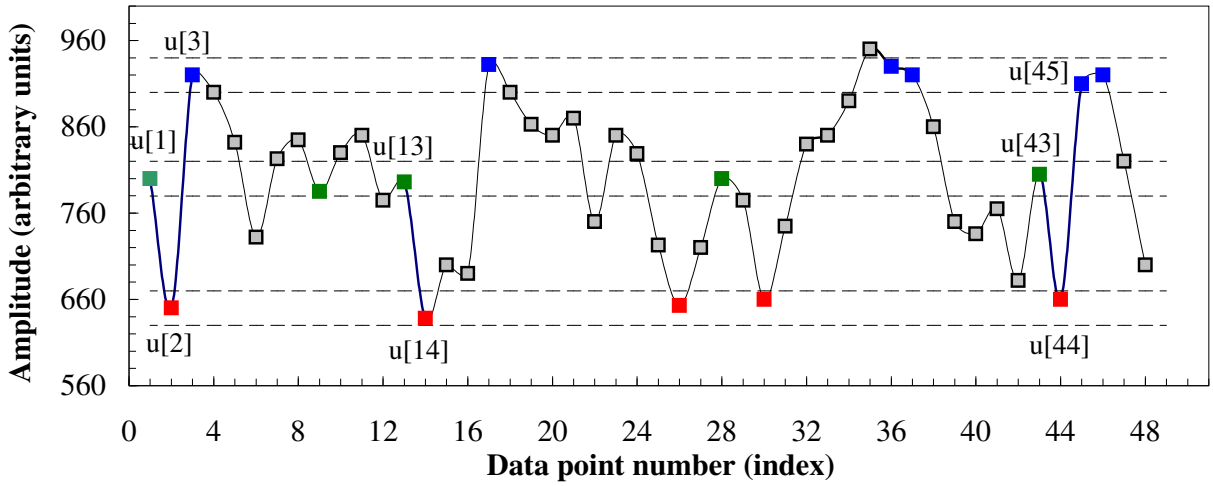


Figure 2: A simulated time series u[1], . . . , u[n] is shown to illustrate the procedure for calculating sample entropy (SampEn) for the case in which the pattern length, *m*, is 2, and the similarity criterion, *r*, is 20. (*r* is a given positive real value that is typically chosen to be between 10% and 20% of the sample deviation of the time series.) Dotted horizontal lines around data points u[1], u[2] and u[3] represent u[1] $\pm$ *r*, u[2] $\pm$ *r*, and u[3] $\pm$ *r*, respectively. Two data values match each other, that is, they are indistinguishable, if the absolute difference between them is $\leq$ *r*. All green points represent data points that match the data point u[1]. Similarly, all red and blue points match the data points u[2] and u[3], respectively. Consider the 2-component green-red template sequence (u[1], u[2]) and the 3-component green-red-blue (u[1], u[2], u[3]) template sequence. For the segment shown, there are two green-red sequences, (u[13], u[14]) and (u[43], u[44]), that match the template sequence (u[1], u[2]) but only one green-red-blue sequence that matches the template sequence (u[1], u[2], u[3]). Therefore, in this case, the number of sequences matching the 2-component template sequences is two and the number of sequences matching the 3-component template sequence is 1. These calculations are repeated for the next 2-component and 3-component template sequence, which are, (u[2], u[3]) and (u[2], u[3], u[4]), respectively. The numbers of sequences that match each of the 2- and 3-component template sequences are again counted and added to the previous values. This procedure is then repeated for all other possible template sequences, (u[3], u[4], u[5]), . . . , (u[N-2], u[N-1], u[N]), to determine the ratio between the total number of 2-component template matches and the total number of 3-component template matches. SampEn is the natural logarithm of this ratio and reflects the probability that sequences that match each other for the first two data points will also match for the next point.

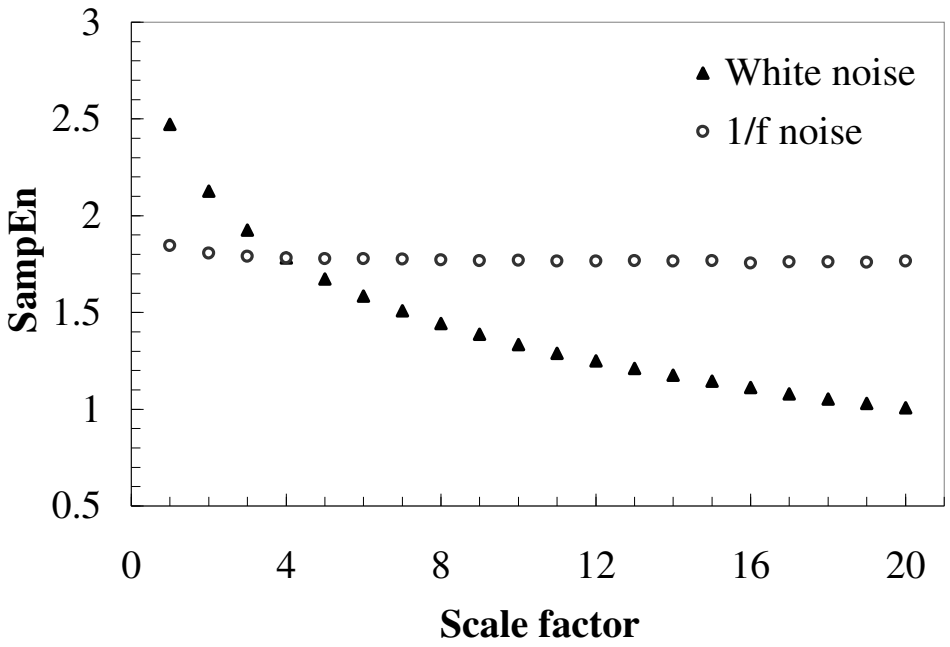


Figure 3: MSE analysis of simulated white and 1/f noise time series. Symbols represent mean values over 30 time series. Parameters to calculate sample entropy are: $m = 2$, $r = 0.15$, and $N = 30{,}000$. Adapted from reference [2].

```
ann2rr -r RECORD -a ANNOTATOR -A -i s4 >RECORD.rr
```

where `RECORD` is the record name and `ANNOTATOR` is the annotator name of the beat annotation file you wish to study. (If you choose a PhysioBank record and have not previously downloaded the annotation file into a local directory, ann2rr obtains the annotations directly from PhysioNet. For details on the options used in this command, see ann2rr in the WFDB Applications Guide.) For example, the command line:

```
ann2rr -r nsr2db/nsr040 -a ecg -A -i s4 >nsr040.rr
```

creates an interval list from the `ecg` beat annotations for record `nsr040` of the Normal Sinus Rhythm RR Interval Database. The first few lines of output from this command are:

```
0.8984
0.7109
0.7188
0.7188
0.7109
0.7031
0.7031
0.7031
0.7031
 ...
```

Of course, `mse` can accept text files containing any similarly formatted series; it is not restricted to use with RR interval time series.

## Setting parameters for `mse`

In order to calculate entropy, the values of parameters $m$ and $r$ defining the pattern length and the similarity criterion [3] respectively, have to be fixed. The default values for these parameters are $m = 2$ and $r = 0.15$. The options **-m** and **-r** may be used to change the default values. It is possible to run MSE for a set of different $m$ and $r$ values using the options **-M**, **-b**, **-R**

and **-c**, which specify, respectively, the maximum *m* value, the difference between consecutive *m* values, the maximum *r* value, and the difference between consecutive *r* values. For example, the command line:

```
mse -m 2 -M 4 -b 1 -r 0.15 -R 0.2 -c 0.01 <nsr040.rr >nsr040.mse
```

calculates the MSE curves for the file `nsr040.rr` for all combinations of *m* (2, 3, and 4) and *r* (0.15, 0.16, 0.17, ..., 0.2) values.

The entire time series or a segment may be selected for the MSE analysis. By default the first 40,000 data points (or the entire time series if it contains fewer than 40,000 data points) are selected. The user may select a different segment using the options **-i** and **-I** that specify the first and the last points of the segment.

The MSE curve is calculated for a range of scales, typically from 1 to 20 data points. Each scale defines the length of the window used for building the coarse-grained time series. The user may change the maximum scale value and the difference between consecutive scale values using the options **-n** and **-a** respectively. For example, if we run the command line:

```
mse -n 10 -a 2 <nsr040.rr >mse-nsr040
```

we obtain an output file (`MSE-nsr040`) containing:

```
m = 2,  r = 0.150

1       0.235
3       0.192
5       0.238
7       0.257
9       0.277
```

In this output, the first column contains the scale factors, and the second column provides the corresponding entropy values.

Several time series may be analyzed simultaneously with the option **-F**. For this purpose a list with the names of the data files (one per line) should be saved as a text file. For example, we can generate an RR interval series for record nsr047 following the same method as for nsr040, above, and then create a file named `filelist`, containing:

```
nsr040.rr
nsr047.rr
```

We can then process both of these files using `mse` by the command:

```
mse -n 10 -a 2 -F filelist >filelist.mse
```

to obtain this output:

```
m = 2,  r = 0.150

        nsr040  nsr047
1       0.235   0.796
3       0.192   1.053
5       0.238   1.218
7       0.257   1.228
9       0.277   1.201

*************************
Mean and SD over all files
*************************

        m=2, r=0.150
        mean    sd
1       0.515   0.397
3       0.623   0.609
5       0.728   0.693
7       0.742   0.686
9       0.739   0.654
```

For each scale, the mean and the sample deviation of the entropy values over all data files are calculated.

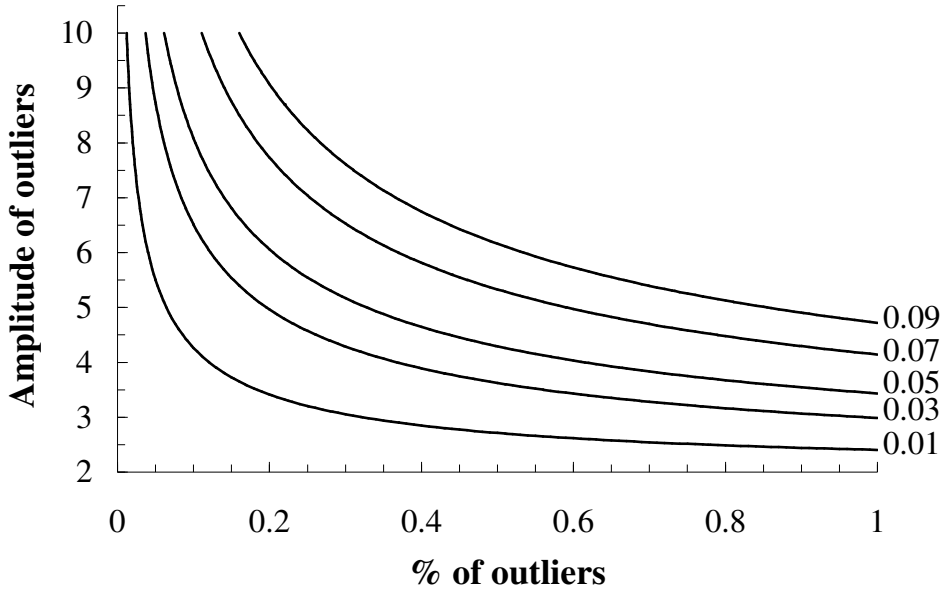


Figure 4: Contour plot showing how the percentage of outliers and their amplitude (relative to the mean value of the time series) affects the variance of the time series. Lines connect $(x, y)$ pairs of values that change the variance by the same amount.

### Summary of options and default values for `mse`

**Scale factor**

**-n**: maximum scale factor (default: 20)
**-a**: difference between consecutive scale factors (default: 1)

**Parameter *m***

**-m**: minimum m value (default: 2)
**-M**: maximum m value (default: 2)
**-b**: difference between consecutive m values (default: 1)

**Parameter *r***

**-r**: minimum r value (default: 0.15)
**-R**: maximum r value (default: 0.15)
**-c**: difference between consecutive r values (default: 0.05)

**Segment selection**

**-i**: starting data point (default: 0)
**-I**: ending data point (default: 39999)

**Multiple data files**

**-F**: text file; each line lists the name of a data file

## 4 Effect of outliers on the MSE curves

Outliers may affect the entropy values because they change the time series standard deviation and therefore, the value of the parameter *r* that defines the similarity criterion. Figure 4 shows that a small number of outliers with high amplitude has similar effects on the variance as a higher percentage of outliers with lower amplitude.

Figure 5 presents three RR interval time series derived from a 24 hour Holter recording of a healthy subject (nsr020). We calculate the MSE curves for a segment of the original time series (file 1) and two filtered time series (file 2 and file 3). File 1 contains the first 30,000 data points (RR intervals) of the original time series. File 2 contains the same data as file 1, but

excluding the 6 RR intervals that exceed 2s. Similarly, file 3 contains these same intervals, but excluding the 43 RR intervals that are less than 0.3s or greater than 1s.

Using `mse`, we can obtain the following MSE analysis of these three files:

| Scale | File 1 | File 2 | File 3 |
|---|---|---|---|
| 1 | 0.009 | 0.734 | 0.734 |
| 3 | 0.012 | 0.937 | 0.933 |
| 5 | 0.012 | 1.137 | 1.140 |
| 7 | 0.011 | 1.138 | 1.144 |
| 9 | 0.011 | 1.222 | 1.210 |
| 11 | 0.011 | 1.174 | 1.224 |
| 13 | 0.011 | 1.204 | 1.186 |
| 15 | 0.011 | 1.199 | 1.186 |
| 17 | 0.010 | 1.183 | 1.189 |
| 19 | 0.009 | 1.186 | 1.212 |

File 1 includes 6 outliers (225.8, 4.43, 5.24, 4.65, 4.40, 8.61) at least one order of magnitude higher than the mean value of the time series. The sample deviations of the contents of files 1, 2 and 3 are 1.3, 0.62 and 0.60, respectively. For file 1, any two data points $x_i$ and $x_j$ such that $|x_i - x_j| \leq 0.2$s are not distinguishable. Therefore, this time series seems to be very regular and the entropy values are close to zero. File 3 contains 37 fewer outliers than file 2. However, since the difference between their sample deviations is less than 0.05%, the entropy values are very close. We note that the inclusion of a low percentage of outliers does not significantly affect MSE analysis unless their differences from the mean value of the time series are orders of magnitude larger than the sample deviation.

# References


[1] Costa M., Goldberger A.L., Peng C.-K. Multiscale entropy analysis of biological signals. *Phys Rev E* 2005;**71**:021906.

[2] Costa M., Goldberger A.L., Peng C.-K. Multiscale entropy analysis of physiologic time series. *Phys Rev Lett* 2002;**89**:062102.

[3] Richman J.S., Moorman J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am J Physiol Heart Circ Physiol* 2000;**278**(6):H2039-H2049.

[4] Pincus S.M. Approximate entropy as a measure of system complexity. *Proc Natl Acad Sci USA* 1991;**88**:2297-2301.

[5] Lake D.E., Richman J.S., Griffin M.P., Moorman J.R. Sample entropy analysis of neonatal heart rate variability. *Am J Physiol Regul Integr Comp Physiol* 2002;**283**(3):R789-97.

[6] Pincus S.M. Assessing serial irregularity and its implications for health. *Ann N Y Acad Sci* 2002;**954**:245-67.

[7] Grassberger P. Information and complexity measures in dynamical systems, in Atmanspacher H, and Scheingraber H (eds.), *Information Dynamics*. New York: Plenum Press, 1991; 15-33.

[8] Costa M., Goldberger A.L., Peng C.-K. Multiscale entropy to distinguish between physiologic and synthetic RR time series. *Computers in Cardiology* 2002;**29**:137-140.

[9] Costa M, Peng C-K, Goldberger AL, Hausdorff JM. Multiscale entropy analysis of human gait dynamics. *Physica A* 2003;**330**:53-60.

[10] Costa M., Healey J.A. Multiscale entropy analysis of complex heart rate dynamics: discrimination of age and heart failure effects. *Computers in Cardiology* 2003;**30**:705-708.

[11] Nikulin V.V., Brismar T. Comment on "Multiscale entropy analysis of complex physiologic time series". *Phys Rev Lett* 2004;**92**(8):089803.

[12] Costa M., Goldberger A.L., Peng C.-K. Reply. *Phys Rev Lett* 2004;**92**(8):89804.

[13] Zhang Y.C. Complexity and 1/f noise: a phase space approach. *J Phys I France* 1991;**1**:971-977.

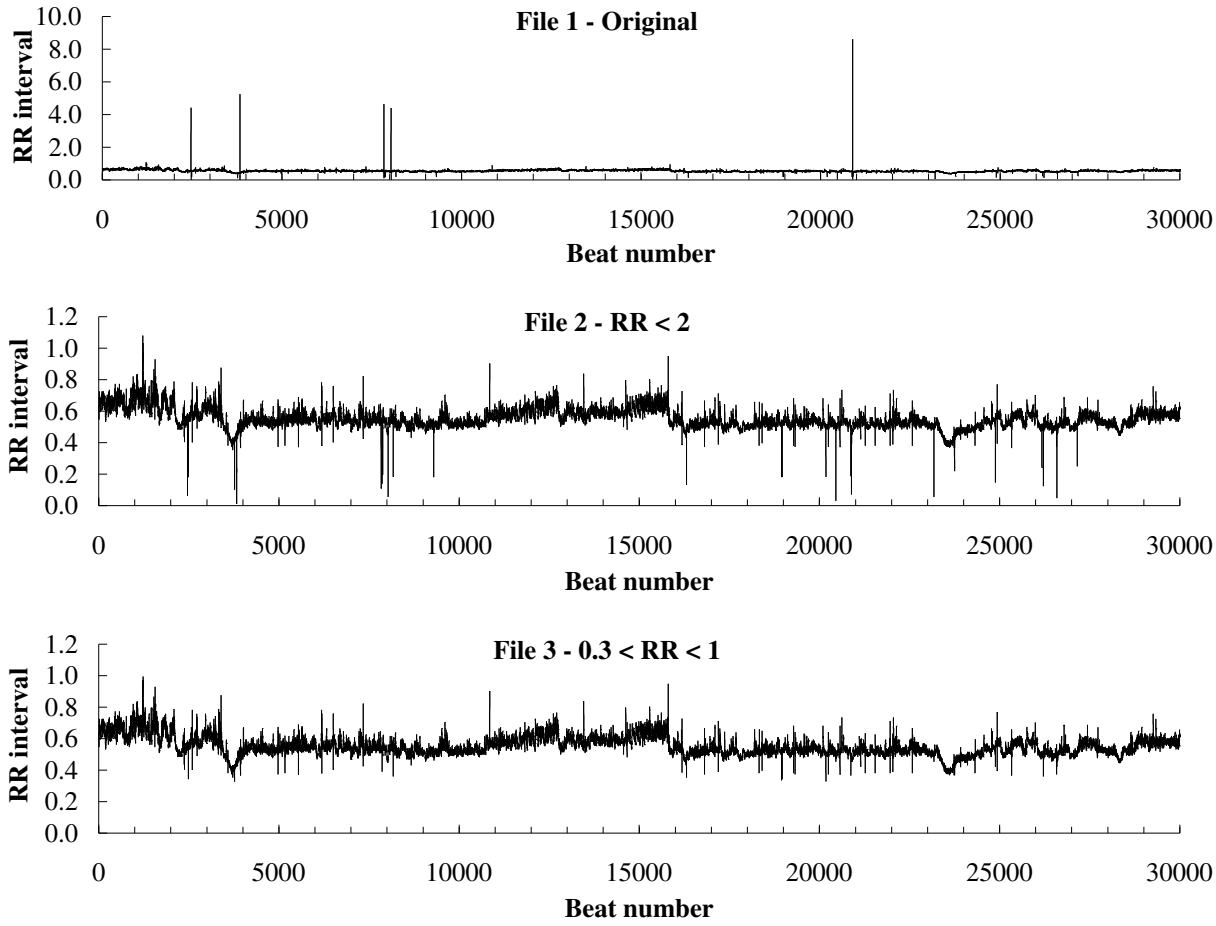


Figure 5: Top panel: cardiac interbeat (RR) interval time series from a healthy subject (nsr020). One outlier (225.8) is not represented. Middle panel: Time series obtained from the time series presented in the top panel excluding the 6 RR intervals $< 2$s. Bottom panel: Time series obtained from the time series presented in the top panel excluding the 43 RR intervals outside the interval 0.3 to 1s.