TIPNet
(Temporal Information Partitioning Network)
User Guide for MATLAB toolbox

Allison Goodwell

April 28, 2017

1 Introduction

This Matlab interface takes inputs of time-series datasets as "nodes" in a network, and computes information measures to identify and characterize time dependencies between nodes.

1.1 Quick Start

Run the file called EntropyGUI_mainwindow.m. Click **Load New Data** option, and load either a .mat or .xls file containing columns of time series data. A .mat project file that has been loaded and saved previously can also be loaded for immediate viewing of results in the **Load Project file** option. For a .xls file, variable names should be the top row of the file and the first column should be a time step. A .mat file should include a (# variables x # timesteps) matrix called "data" and a (1 x # variables) cell called *varnames* with variable names. For any pre-processing, *pdf* options, or network computation options, see the appropriate section. To compute a single network using all data with default options, click on **Compute Links**. All results are stored in the *entropy* structure that is saved in the project file. Results can then be viewed by clicking Plot Results. The results are saved in a structure called **entropy**, the contents of which are described in this manual and in the included excel file called **TIPNet_nomenclature.xls**.

2 Information Measures

2.1 Entropy and Mutual Information

$$H(X) = -\sum p(x)\log_2(p(x)) \tag{1}$$

$$I(X;Y) = H(Y) - H(Y|X) = \sum p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{2}$$

where X and Y are time-series variables that may be simultaneous or involve some time lag between them. When we consider X to be a "source" node and Y to be a "target" node, the quantity $I(X;Y)$ indicates the strength of a link from

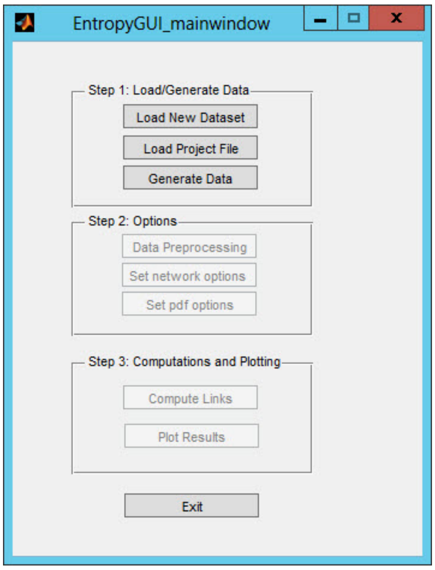


Figure 1: Main screen, one of first 3 buttons must be chosen to load data or project file, or generate test data.

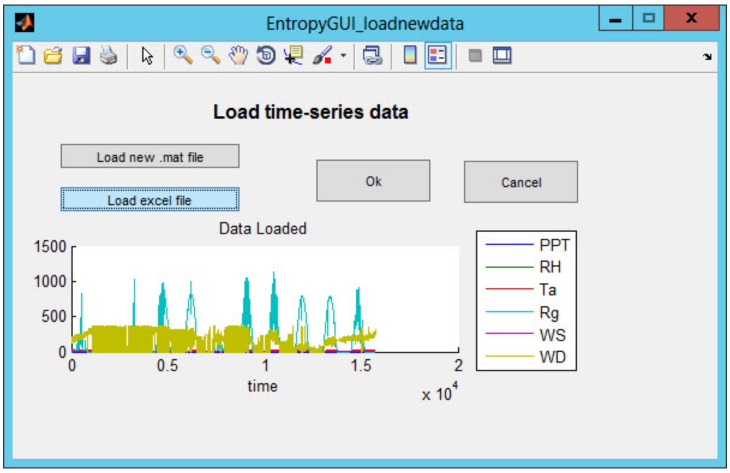


Figure 2: Example weather station data set loaded from excel.

X to Y in that X reduces the uncertainty of the Y. For a range of lag times τ, $I(X(t-\tau);Y)$ is computed. Transfer Entropy $T_E(X(t-\tau) \rightarrow Y)$, which is a special case of conditional information $I(X(t-\tau);Y|Y(t-1))$ is also computed as follows:

$$T_E(X \rightarrow Y) = I(X;Y|Y_1) = \sum_{x,y,y_\tau} p(x,y,y_1) \log \left[\frac{p(x,y,y_1)}{p(y,y_1)} \right] \tag{3}$$

where abbreviated symbols are $x = x(t-\tau)$, $y = y(t)$, and $y_1 = y(t-1)$.

As discussed in [2], T_E omits a redundant component (overlapping information shared to target $Y(t)$ by both $X(t-\tau)$ and $Y(t-1)$) but adds in a synergistic component (information shared to the target $Y(t)$ due to knowledge of both sources together).

The dominant time scale of the link from X to Y is the $\tau > 0$ corresponding either to the maximum $I(X(t-\tau);Y)$ (bits) or the normalized value $\frac{I(X(t-\tau);Y)}{min(H(X),H(Y))}$ (bits/bit), depending on the **mi.NormOpt** parameter (see next section).

2.2 PDF estimation and statistical significance

Computation of these measures involves estimating joint probability density functions (*pdf*) for lagged X and Y. This program includes a fixed bin method [6, 5] or a Kernel Density Estimation method [5, 8] to estimate *pdf*s from data. While the fixed binning method tends to be faster, the *KDE* method can be advantageous for sparse data sets since it smooths the *pdf* based on the sample size. For any detected $I(X;Y)$ value, we test for statistical significance using a shuffled-surrogate hypothesis test in which the time-series data are shuffled randomly to destroy any time correlations. Mutual information is then computed for $N = 100$ (default) surrogates of shuffled data, and a 99% significance test is performed to assess whether the computed measure is significantly stronger than links detected from the shuffled surrogates [2, 6].

2.3 Information Partitioning Measures

Once links are detected based on lagged mutual information, we further assess each link in terms of its uniqueness, synergy, or redundancy by analyzing its relationship with other links to the same target. As introduced in [9] and discussed in [2, 1, 3, 4], the total information shared between 2 source nodes X_1 and X_2 to a target Y can be partitioned into four components as follows:

$$I(X_1, X_2; Y) = U_1(Y; X_1) + U_2(Y; X_2) + R(Y; X_1, X_2) + S(Y; X_1, X_2) \tag{4}$$

where U_1, U_2, R, and S are non-negative quantities. R is information that both sources share with the target *redundantly*, U_1 and U_2 are information that only X_1 and X_2, respectively share with the target *uniquely*, and S is information that is provided to the target only when both sources are known together, or *synergistically*. Individual mutual information terms decompose as [9]:

$$I(Y; X_1) = U_1 + R \tag{5}$$
$$I(Y; X_2) = U_2 + R. \tag{6}$$

The proposed redundancy measure R_{MMI} [9, 1] is actually an upper bound for redundant information:

$$R_{MMI} = \min[I(X_1; Y), I(X_2; Y)] \tag{7}$$

The minimum bound of redundant is as follows [3]:

$$R_{\min} = \max[0, I(X_1;Y) + I(X_2;Y) - I(X_1,X_2;Y)] \tag{8}$$

We implement a scaled version of R:

$$R = R_{\min} + I_s(R_{MMI} - R_{\min}) \tag{9}$$

where $I_s =$ is the scaled source dependency, so that independent sources X_1 and X_2 result in minimum redundancy and highly dependent sources result in maximum redundancy. After R is computed for a given two sources to a target, the quantities U_1, U_2, and S can be computed directly. For a network of multiple interacting nodes, we consider each pair of sources to a target and evaluate the redundancy, uniqueness and synergy of each source pair.

We define a measure T/I as follows [2]:

$$\frac{T}{I}(X_{s1}|X_{s2} \to X_{tar}) = \frac{U_{s1} + S_{s1,s2}}{U_{s1} + U_{s2} + S_{s1,s2} + R_{s1,s2}} = \frac{I(X_{tar};X_{s1}|X_{s2})}{I(X_{tar};X_{s1},X_{s2})} \tag{10}$$

For each source link X_{s1}, we define $T/I(X_{s1} \to X_{tar})$ as the minimum value of Equation (10) given any other source node X_{s2} as follows:

$$\frac{T}{I}(X_{s1} \to X_{tar}) = \min_{X_{s2}}\left[\frac{T}{I}(X_{s1}|X_{s2} \to X_{tar})\right] \tag{11}$$

In this way, T/I for a source to target link indicates the relative uniqueness and synergy of that link with respect to each other source to the same target.

We apply Equation 9 to compute U_1, U_2, R, and S components for every pair of sources to a target. Similarly to T/I, we define the components for each link as follows:

$$R(X_{s1} \to X_{tar}) = \max_{X_{s2}}[R(X_{s1},X_{s2};X_{tar})] \tag{12}$$

$$U(X_{s1} \to X_{tar}) = \min_{X_{s2}}[U(X_{s1},X_{s2};X_{tar})] \tag{13}$$

$$S(X_{s1} \to X_{tar}) = \max_{X_{s2}}[S(X_{s1},X_{s2};X_{tar})] \tag{14}$$

The matrices $R(X,Y)$ and $R_pair(X,Y) = Z$ in the *entropy* results structure identifies the redundancy (in bits) that source node X shares with target Y along with source node Z. The source nodes X and Z are the most highly redundant sources to Y. Similarly, $S(X,Y)$ and $S_pair(X,Y) = W$ in the *entropy* results structure identifies the synergistic information (in bits) that source node X shares with target Y along with source node W, and X and W are the most strongly synergistic sources to Y.

The matrices R, S, and U are used for plotting purposes in this program, but we also compute and retain the measures for each pair of source variables in the results structure *entropy* as S_allpairs, R_allpairs, and U_allpairs.

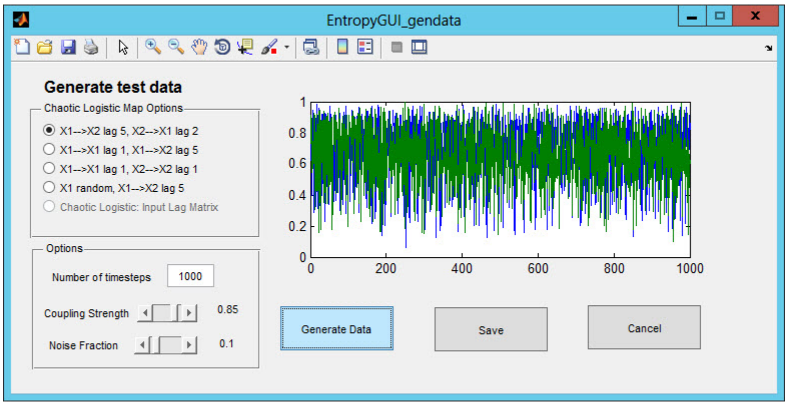
3 Guide

3.1 Getting Started

Important! First Time Use Only If you choose to use the KDE method for *pdf* computations, you must compile 3 C-mex files in MATLAB as follows: Go the the Functions folder, then type in the command line *mex -mdKDE_1d.c*. If an error occurs, you may need to choose a C compiler. Perform the same operation for *mdKDE_2d.c* and *mdKDE_3D.c*. This only needs to be done the first time you use the program.

Run the file called *EntropyGUI_mainwindow.m*. Click **Load New Data** option, and load either a .mat or .xls file containing columns of numeric time series data. Examples of .mat files and .xls files containing weather station data are provided in the folder projects_datasets. For a .xls file, variable names should be the top row of the file. A .mat file must include a (# variables x # timesteps) matrix called *data* and a (1 x # variables) cell called *varnames* with variable names. Once a data set is loaded, click **OK** to save the file as a project file. This project file will contain the **mi** (**m**odel **i**nformation) structure with all default parameters to run the temporal network program. When parameters are altered in the **pre-processing**, **network option**, or **pdf options**, they are updated in the **mi** structure in the project file. To reset all parameters to their default values, load the data as a new data set. To re-load a project file with any parameters that have been previously altered from default values, choose the load project option on the main screen.

3.2 Generating Test Data



● X1-->X2 lag 5, X2-->X1 lag 2
○ X1-->X1 lag 1, X1-->X2 lag 5
○ X1-->X1 lag 1, X2-->X2 lag 1
○ X1 random, X1-->X2 lag 5
○ Chaotic Logistic: Input Lag Matrix

Number of timesteps 1000
Coupling Strength 0.85
Noise Fraction 0.1

Figure 3: Example generated chaotic logistic test data. User can choose 2-node cases with several forcing options, and change the coupling strength between nodes and the random noise component.

Alternatively to loading a time series data set, the **Generate Data** option generates a 2-node chaotic logistic time series data set for one of four different forcing cases:

1. Feedback forcing, where $X1$ and $X2$ drive each other:

$$X2(t) = 4X1(t-5)[1 - X1(t-5)] \tag{15}$$
$$X1(t) = 4X2(t-2)[1 - X2(t-2)] \tag{16}$$

2. $X1$ drives itself via the chaotic logistic equation and also drives $X2$:

$$X2(t) = 4X1(t-5)[1 - X1(t-5)] \tag{17}$$
$$X1(t) = 4X1(t-1)[1 - X1(t-1)] \tag{18}$$

3. $X1$ and $X2$ are independent, each driven by the chaotic logistic equation:

$$X2(t) = 4X2(t-1)[1 - X2(t-1)] \tag{19}$$
$$X1(t) = 4X1(t-1)[1 - X1(t-1)] \tag{20}$$

4. $X1$ is a uniform random variable, and drives $X2$ through the chaotic logistic equation:

$$X2(t) = 4X1(t-5)[1 - X1(t-5)] \tag{21}$$
$$X1(t) = U(0,1). \tag{22}$$

For any case, the noise fraction slider bar for $0 \leq \epsilon_z \leq 1$ can be altered to add a degree of randomness into every node. For example, $\epsilon_z = 1$ generates 2 independent uniform random nodes, regardless of the chosen case.

3.3 Options

After loading a project file, new data file, or generated test data, there are three buttons to alter network parameters and properties from default values. These options include *pdf* estimation methods, network run options, and time-series pre-processing.

3.3.1 Pre-Processing Options

Each timeseries variable X is normalized between (0,1) as follows:

$$X_norm = \frac{(X - X_{min})}{X_{max} - X_{min}} \tag{23}$$

This is a default option for *pdf* computations, but users can select and unselect this option in the pre-processing screen only for viewing purposes.

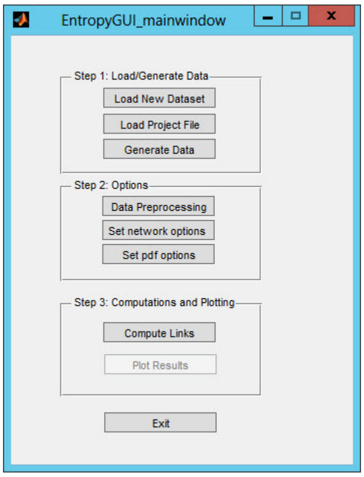
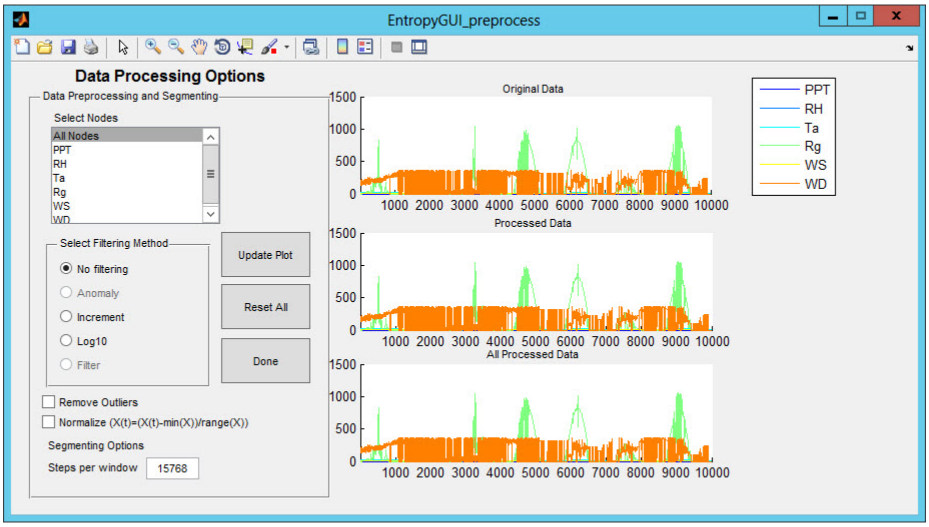


Figure 4: Main TIPNet screen after a project file, new data file, or generated data has been chosen.



● No filtering
○ Anomaly
○ Increment
○ Log10
○ Filter

□ Remove Outliers
□ Normalize (X(t)=(X(t)-min(X))/range(X))

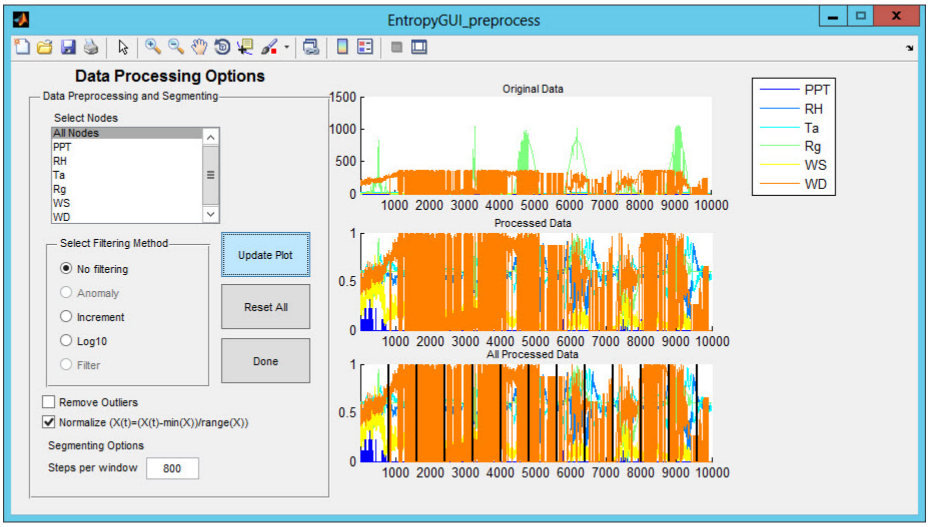
Steps per window 15768

Figure 5: Data preprocessing screen. Individual nodes or all nodes together can be selected and altered, and segments can be determined for separate network time windows.

We include 5 options for data filtering or altering. For each type, there is an option to remove or not remove outliers. Alternately, any processing or filtering can be performed prior to loading data in this program. The types of pre-processing presented here are only basic examples of filtering or outlier removal techniques.

No Filtering This option reverts the data to the original normalized data set.

Anomaly For data that exhibit diurnal or seasonal cycle, the X-day anomaly



EntropyGUI_preprocess

Data Processing Options

Data Preprocessing and Segmenting

Select Nodes

All Nodes
PPT
RH
Ta
Rg
WS
WD

Select Filtering Method

● No filtering
○ Anomaly
○ Increment
○ Log10
○ Filter

□ Remove Outliers

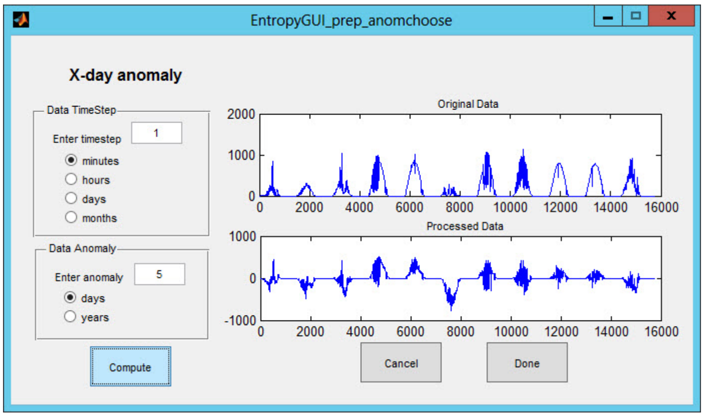
☑ Normalize (X(t)=(X(t)-min(X))/range(X))

Segmenting Options

Steps per window 800

Figure 6: Data preprocessing screen where dataset has been segmented into equal length time windows.

is the difference between the value at a certain time (e.g. 12:00 noon on Day 100) and the mean value at that time on the X surrounding days (e.g. 12:00 noon on Days 95-105 for a 10-day anomaly). The anomaly can only be computed for 1 variable at a time, and the user must check on the time step and units of the data (minutes, days) and units of the desired anomaly (days, years). The anomaly of the originally loaded data is then again normalized to a (0,1) range.



EntropyGUI_prep_anomchoose

X-day anomaly

Data TimeStep

Enter timestep 1

● minutes
○ hours
○ days
○ months

Data Anomaly

Enter anomaly 5

● days
○ years

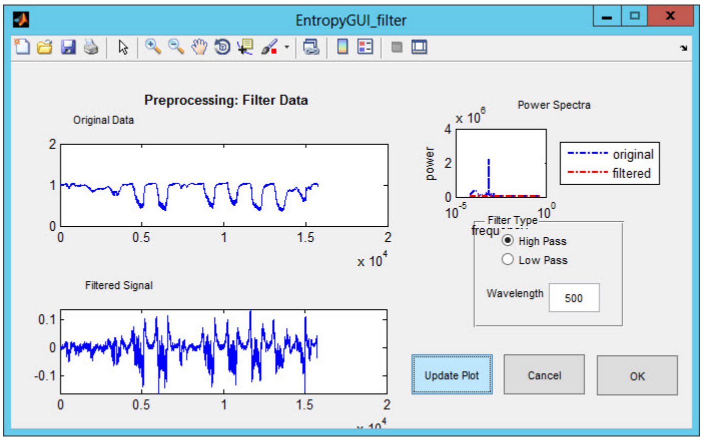
Figure 7: 5-day anomaly of 1-minute resolution shortwave radiation data.

Increment For data where an increase or decrease may be more relevant than an actual value (e.g. a population variable). This changes the data as follows

$$X(t) = X(t) - X(t-1) \tag{24}$$

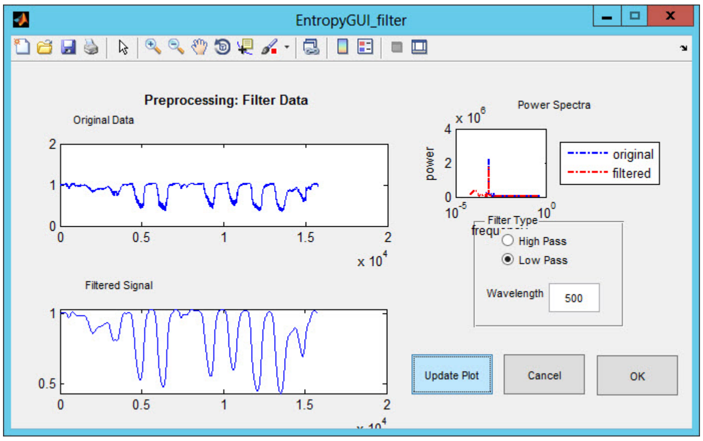
Log 10 : This takes the base 10 logarithm for skewed input data (e.g. flow rate data)

Filter For a single variable at a time, this option applies a Butterworth Filter to the data for a high-pass or low-pass filter to preserve or omit short-term fluctuations. This can be used to *(a)* omit the diurnal and/or seasonal cycle with a high-pass filter *(b)* omit noise with a low-pass filter, as done in [7].



Filter Type
● High Pass
○ Low Pass
Wavelength 500

Figure 8: High pass filter applied to Relative Humidity data to remove diurnal cycle.



Filter Type
○ High Pass
● Low Pass
Wavelength 500

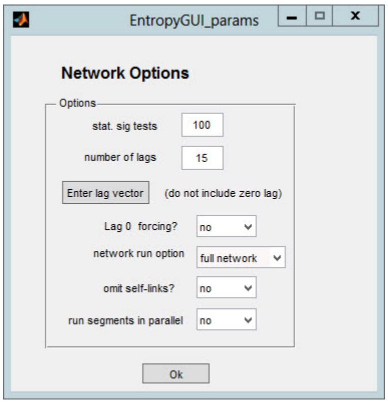
Figure 9: Low pass filter applied to Relative Humidity data to remove noise.

For each option, outlier removal is performed after the operation (e.g. after taking the logarithm or increment). Outliers, data points that lie above $X_{75} +$

$1.5IQR$ or below $X_{25} - 1.5IQR$, are set to the values $X_{75} + 1.5IQR$ or $X_{25} - 1.5IQR$, respectively rather than being removed. Removal of outliers would impact the time dependencies by removing a time-step of the specified variable. Any other outlier removal via gap-filling or other methods should be done prior to loading a dataset.

Finally, to partition a long time-series data sets into multiple segments, the segment length can be changed. This option results in computation of one network for each time-series segments, and is useful to compare before-after scenarios or to consider the evolution over time of interactions.

3.3.2 Network Options



EntropyGUI_params

Network Options

Options

stat. sig tests: 100

number of lags: 15

Enter lag vector (do not include zero lag)

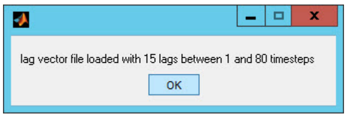
Lag 0 forcing?: no

network run option: full network

omit self-links?: no

run segments in parallel: no

Figure 10: Network Options screen.



lag vector file loaded with 15 lags between 1 and 80 timesteps

Figure 11: Lag vector file (.mat file with vector called lagvect) loaded from the folder **UserData**. Network options screen will automatically populate with number of lags in the file.

The Network Options screen contains several options:

Statistical Sig Tests: The shuffled surrogates method is used to determine statistical significance of each computed $I(X_1; X_2)$ value. The default number of significance tests is 100.

Number of Lags: The number of lags for which lagged information measures are to be computed as $\tau = 1...\text{nlags}$.

Enter Lag Vector: Alternatively to specifying a number of consecutive lags, load a .mat file called lagvect.mat with a vector of lags named lagvect, containing lags. This can be used to compute lags at intervals, for example lagvect = [5 10 15 30 60 120] to compute network measures at only 6 time lags but for different lag times than 1-6. A *lagvect.mat* file is provided in the folder *UserData*, and can be overwritten as needed. The lag vector should consist of non-negative integers, and should not include zero (see next point).

Lag Zero Forcing: By default, zero-lag or instantaneous mutual information is not considered as a dominant link that can be redundant, synergistic, or unique with any other link. To include zero-lag forcing (e.g. if the time step is such that X may be expected to drive Y at a time scale much lower than the time step), change this option to *Yes.*

Network Run Option: By default, the program will perform all computations for mutual information, transfer entropy, and information decomposition as described in the previous section. To only compute individual node entropy or mutual information, change this option as appropriate. Note: the **Plot Results** viewer will not function if this option is altered (i.e. no plotting for any case except full network, however the **entropy** results structure will still be saved with the limited results)

Omit Self-Links: By default, node X is considered as a potential source to itself, and a detected link $I(X(t-\tau);X(t))$ may be unique, synergistic, or redundant when another link to X is considered. To omit these “self” links, change this option to *Yes.*

Run Segments in Parallel: If your data set is segmented into multiple time series in the **Pre-Processing Options** and your computer can run parallel code in Matlab (parfor), enable this to run segments in parallel.

3.4 PDF options

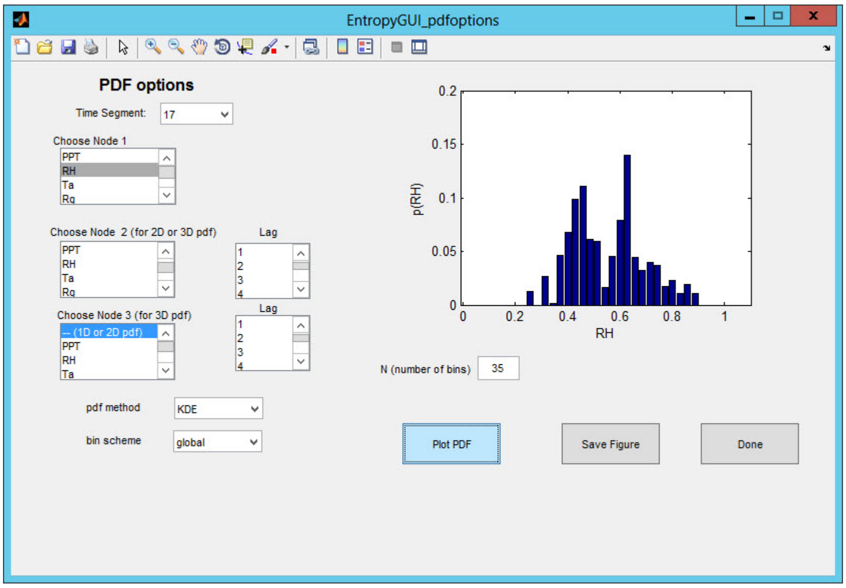
All information measures computed in this program are based on 1D, 2D, and 3D *pdf*s. This screen allows you to view these *pdf*s and alter certain parameters, namely the *pdf* estimation method, number of bins, and local or global binning.

Time Segment For data sets that have been segmented in **Pre-Processing Options**, choose segment to view pdf.

Choose nodes and lags Choose 1,2 or 3 nodes to view 1D, 2D, or 3D *pdf*, respectively. To view lagged *pdf*, choose lag for second and third nodes. A 1D *pdf* will appear as a bar chart where the height of each bar corresponds to $p(x)$. A 2D *pdf* will appear as a color scaled image where the color corresponds to $p(x,y)$. A 3d *pdf* will appear as a 3D point cloud, where a point represents a $p(x,y,z)>0$.

N Number of bins or locations at which to compute *pdf.* The default value is $N=25$, and N can range up to 100.

pdf method Choose between the KDE method and fixed bin method (default). For the KDE method, optimal parameters for the Epanechnikov kernel (based on [8], we refer to [2] for details) are chosen for each time window and each data point.



EntropyGUI_pdfoptions

PDF options

Time Segment: 17

Choose Node 1
PPT
RH
Ta
Rg

Choose Node 2 (for 2D or 3D pdf)
PPT
RH
Ta
Rg

Lag
1
2
3
4

Choose Node 3 (for 3D pdf)
-- (1D or 2D pdf)
PPT
RH
Ta

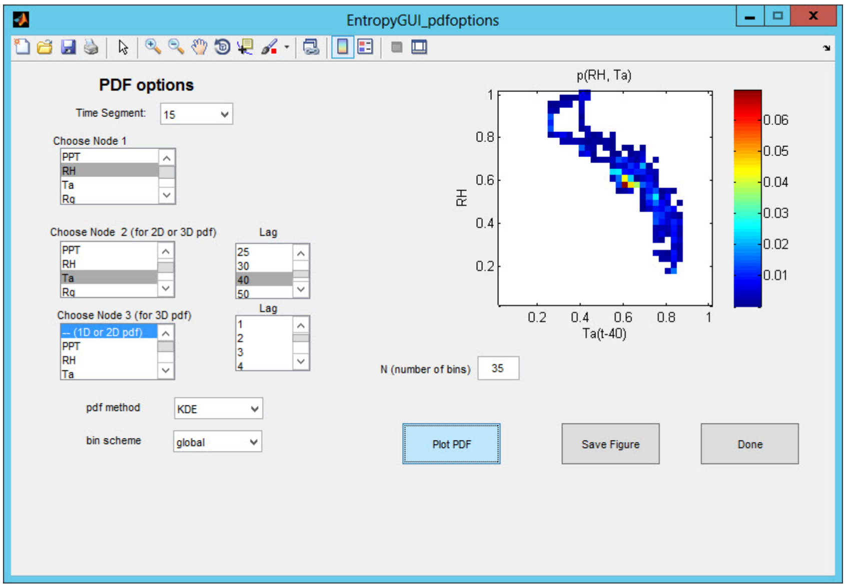
Lag
1
2
3
4

pdf method: KDE

bin scheme: global

N (number of bins): 35

Figure 12: 1d *Pdf* of relative humidity for a specific time segment, using the KDE *pdf* estimation method, global binning, and 35 bins (points at which the *pdf* is estimated).



EntropyGUI_pdfoptions

PDF options

Time Segment: 15

Choose Node 1
PPT
RH
Ta
Rg

Choose Node 2 (for 2D or 3D pdf)
PPT
RH
Ta
Rg

Lag
25
30
40
50

Choose Node 3 (for 3D pdf)
-- (1D or 2D pdf)
PPT
RH
Ta

Lag
1
2
3
4

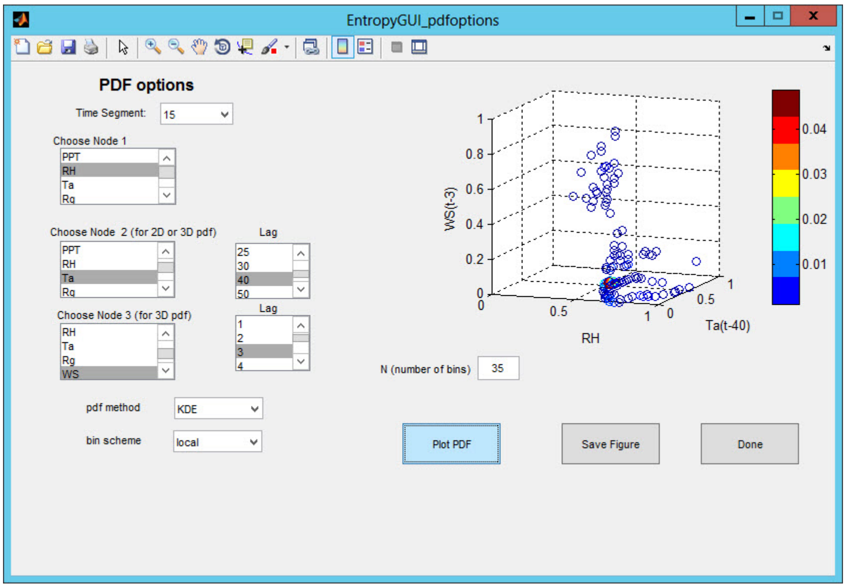
pdf method: KDE

bin scheme: global

N (number of bins): 35

Figure 13: 2d *Pdf* of relative humidity and lagged air temperature for a specific time segment.

bin scheme For segmented data, a global bin scheme (default) scales the data between the global minimum (0) and maximum (1) values. A local bin


EntropyGUI_pdfoptions

PDF options

Time Segment: 15

pdf method: KDE

bin scheme: local

N (number of bins): 35

Figure 14: 3d *Pdf* of relative humidity and lagged temperature and windspeed. Circles in 3d space indicate concentrated areas of *pdf* (i.e. $p(x, y, z) > 0$). In this figure the "local" bin scheme option has been selected so that the particular time window (segment 15 shown here) data is normalized to a 0 to 1 range, rather than global binning.

scheme scales the data for each time window separately between the minimum and maximum values in that segment. For cases with only a single time segment (the default case), local and global binning are the same.

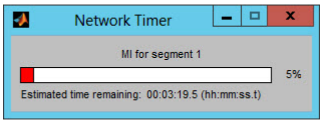
After selecting nodes and/or altering parameters, clicking **Plot PDF** will update the *pdf* plot accordingly. When the KDE method is selected, the **Reset all h values and N** button will reset any previously altered smoothing parameters to the default values and set $N = 25$. Clicking **Done** will save any altered parameters.

3.5 Network Computations and Plotting

Once all options have been selected as desired, click **Compute Links** to construct the temporal information networks.

If the Parallel option is turned off (default option in **Network Options**), a timer window will appear for each segment. For large data sets (typically greater than 1000 data points per segment, more than 20 nodes, or many segments), this could take several minutes to initialize and up to multiple hours to complete. When the Parallel Option is turned on, a progress bar will appear in the Matlab command window. When all computations are finished, the output is saved in the previously created project file in a structure called *entropy*.

Once the output is saved, the **Plot Results** option will be available in the main menu. Click **Plot Results** to view network figures.



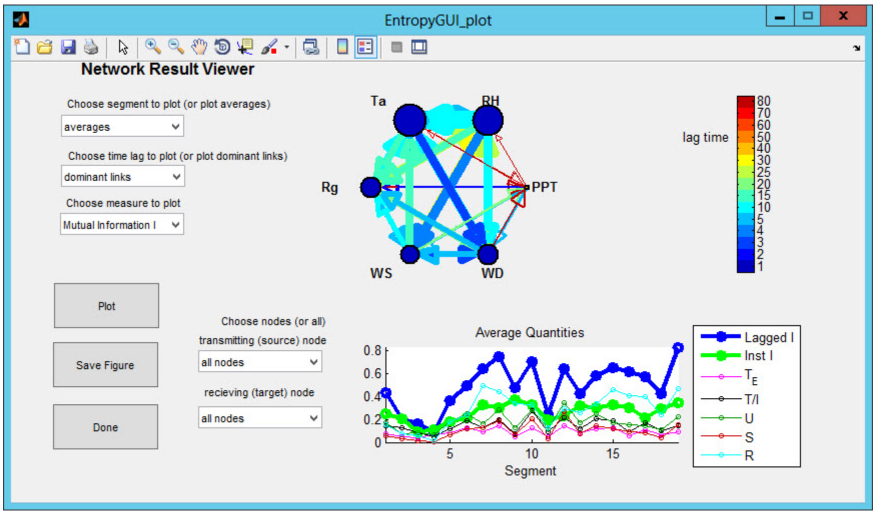
Network Timer

MI for segment 1

5%

Estimated time remaining: 00:03:19.5 (hh:mm:ss.t)

Figure 15: Timer bar will appear for each segment of data set if parallel option is not chosen.



EntropyGUI_plot

Network Result Viewer

Choose segment to plot (or plot averages): averages

Choose time lag to plot (or plot dominant links): dominant links

Choose measure to plot: Mutual Information I

Choose nodes (or all)

transmitting (source) node: all nodes

recieving (target) node: all nodes

Figure 16: Result Viewer for average mutual information links over all time windows. The bottom graph shows overall network statistics for each individual time window.

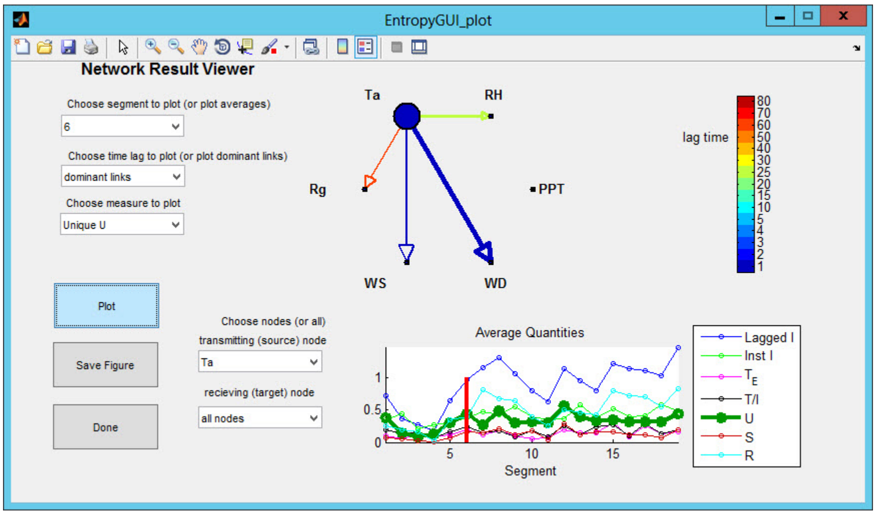
The network circle plots contain each node and depict several information measures and associated time lags and strengths. The arrow indicates directionality (lagged source to target), the color indicates time lag of detected link, and the line width indicates the strength of the link. The node size and color correspond to the “self”-link properties, which may or may not be relevant depending on the selection of **Omit Self Links** in the Network Options. The time series or point plot below the circle network shows each segment (for 1 or more segments) and the total values (averages) for 6 information measures.

Choose Segment This list box is only visible if the data set has been partitioned into multiple segments in the **Pre-Processing Options**.

Choose Time Lag For lagged mutual information and transfer entropy only, values can be plotted for individual values of τ as defined in the lag vector ($mi.lagvect$). For all other measures, only the dominantly detected lags are shown in the circle network plot.

Choose Measure six measures can be plotted as described in the previous section: mutual information, transfer entropy, T/I, U, R, and S

Choose nodes (or all) Select a specific node pair to view only statistics for



EntropyGUI_plot

Network Result Viewer

Choose segment to plot (or plot averages): 6

Choose time lag to plot (or plot dominant links): dominant links

Choose measure to plot: Unique U

Choose nodes (or all)

transmitting (source) node: Ta

recieving (target) node: all nodes

Figure 17: Result Viewer, showing unique information U for a single node (air temperature) as a source node, for a single time window (segment 6, as indicated by the red vertical line in the bottom graph of time windows.

that link, or a single source or target node to view out-going or incoming links, respectively.

References

[1] Adam B. Barrett. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Physical Review E*, 91(5), 2015.

[2] Allison Goodwell and Praveen Kumar. Information Theoretic Measures to Infer Feedback Dynamics in Coupled Logistic Networks. *Entropy*, 17(11):7468–7492, 2015.

[3] Allison Goodwell and Praveen Kumar. Temporal information partitioning: Characterizing synergy, uniqueness, and redundancy in interacting environmental variables. *in review at Water Resources Research*, 2017.

[4] Allison Goodwell and Praveen Kumar. Temporal information partitioning networks (tipnets): A process network approach to infer ecohydrologic shifts. *in review at Water Resources Research*, 2017.

[5] Joon Lee, Shamim Nemati, Ikaro Silva, Bradley A. Edwards, James P. Butler, and Atul Malhotra. Transfer entropy estimation and directional coupling change detection in biomedical time series. *Biomedical Engineering Online*, 11, APR 13 2012.

[6] Benjamin L. Ruddell and Praveen Kumar. Ecohydrologic process networks: 1. identification. *Water Resources Research*, 45, MAR 25 2009.

[7] Alicia Sendrowski and Paola Passalacqua. Process connectivity in a naturally prograding river delta. *Water Resources Research*, 53, 2017.

[8] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[9] Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.