GENERAL INSTRUCTIONS FOR

THE ISMLR PACKAGE

VERSION 1.0

Junjie Zhu
Illinois Institute of Technology
Initially published: August 1, 2017
Last revised: July 23, 2018

UPDATES

ISMLR application		
Version	**Published date**	**Major updates and revisions**
1.0	08/01/2017	• Implemented basic ISMLR method
1.1	01/18/2018	• Combined 1st regressor and 2nd regressor together to avoid confusion • Changed time format "yyyy/mm/dd HH:MM" to a format that depends on input information from GUI and Excel • Revised scripts to show the name of the response regressor in GUI plots and Excel • Downsized fonts in GUI plots
1.2	03/22/2018	• Included "Future information prediction tool"

ISMLR package (includes ISMLR application and data preprocessing tools)		
Version	**Published date**	**Major updates and revisions**

1.0	07/11/2018	• Built based on the ISMLR application ver. 1.2
		• Include data preprocessing functions
		  • outlier detection,
		  • periodogram analysis, and
		  • time-series dataset generation

Table of Content

Page

PREFACE i
UPDATES ii
TABLE OF CONTENT iii
1. INTRODUCTION 1
2. APPLICATION REQUIREMENTS 1
3. OUTLIER DETECTION AND ANALYSIS 1
   3.1 INPUT DATASET REQUIREMENTS 2
   3.2 TECHNICAL OUTLIER ANALYSIS 2
   3.3 STATISTICAL OUTLIER ANALYSIS 4
4. PERIODOGRAM ANALYSIS 8
5. DATASET DEVELOPMENT 9
6. ISMLR APPLICATION 11
   6.1 INPUT DATASET 11
   6.2 SELECTING MODEL PARAMETERS 12
   6.3 COMPUTATION TIME 13
   6.4 OUTPUT INFORMATION AND DOCUMENTS 14
   6.5 EXAMPLE 1 15
   6.6 FUTURE INFORMATION PREDICTION TOOL 19
   6.7 EXAMPLE 2 20
TIPS 23
USE AND DISTRIBUTION 23
ACKNOWLEDGEMENT 23
REFERENCES 24

1. INTRODUCTION

This document is a general guide describing use of the *ISMLR package version 1.0*. In this package are the ISMLR application (version 1.2), and data preprocessing tools for:

- Outlier detection
- Periodogram analysis, and
- Time-series dataset generation

More information about the conventional ISMLR method can be found in the description by Zhu and Anderson (2016). An application of the ISMLR method integrated with a decision-making technique is described by Zhu et al. (2018).

This guide was developed to facilitate a more efficient and wider use of the ISMLR method. The ISMLR package is designed to process a large time-series dataset to:

- Select a subset of important regressors
- Minimize the amount of missing data from a regression analysis
- Provide prediction of the response regressor, or
- Produce pretreated datasets for a subsequent prediction using a more advanced algorithm

In contrast with an ISMLR application where the user is required to provide clean, formatted time-series datasets before running the application, this ISMLR package is designed to help users develop and predict target information based on a raw time-series dataset. Users can also use the built-in tools to perform data pretreatment without performing regression analysis. Reading this document before running the program will help you to better implement the package.

2. APPLICATION REQUIREMENTS

To run the ISMLR package you need to have MATLAB Compiler Runtime (MCR) installed. Any computer (64-bit operating system) that has MATLAB version R2016b installed can directly run the Portable version "*ISMLR package version 1.0.exe*" (file size ≈ 22 MB). To run this file on a computer that has an earlier version of MATLAB or does not have MATLAB installed, requires installation of MCR, version 2016b. Users can download the MCR from the website: https://www.mathworks.com/products/compiler/mcr.html. Alternatively, users can install both components (MCR and ISMLR package) by running the executable installation-based version.

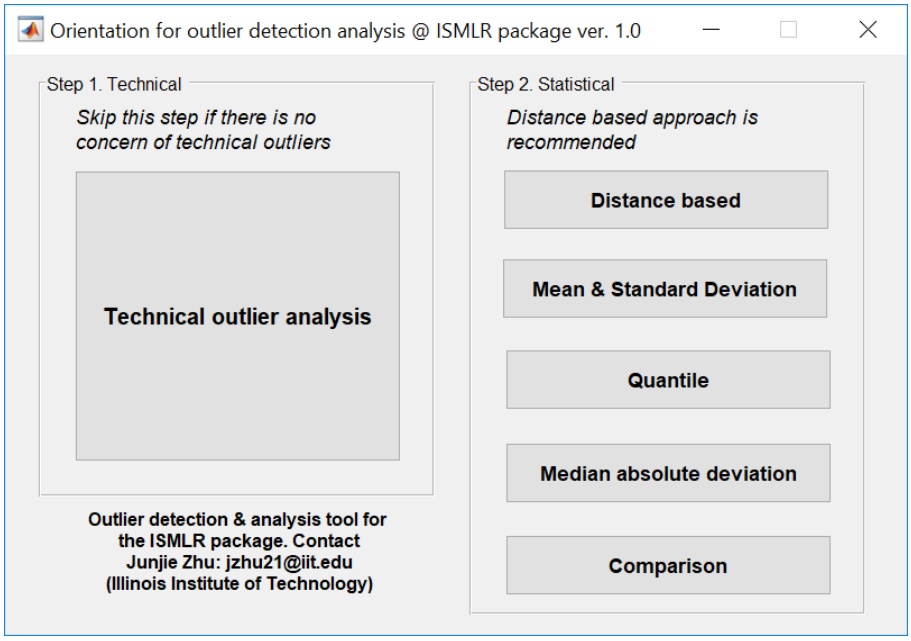
Summary of requirements

- Windows, 64-bit operating system
- MATLAB, version 2016b; or MCR version 2016b (Windows, 64-bit) installed
- Microsoft Excel 2007 or later installed
- Recommended minimum disk space: 2 GB

For more information about "Standalone Applications" or "MATLAB Runtime", refer to the MATLAB® Compiler™ document (MATLAB, 2017).

3. OUTLIER DETECTION AND ANALYSIS

The outlier detection and analysis tool can be used to identify and screen abnormal data from the raw dataset. Two types of outliers, technical and statistical, can be detected. Technical outliers are data that are inconsistent with expert knowledge in the application area. For example, because wastewater treatment plant influent flow is always greater than zero, a reported flow of zero is a technical outlier. Statistical outliers are data that are statistically far away from their sample centers. Typical methods to identify this type of outliers, such as "mean & standard deviation", "quantile", and "median absolute deviation" are included in the package. Also included a "distance based" method, described by Zhu et al. (2015). A "comparison" function can be used to initially compare detection results based on these different methods.



3.1 INPUT DATASET REQUIREMENTS

The input dataset must have one column for the time, and multiple columns for the variables of interest, and the file must be in Excel format (*.xls or *.xlsx). The first row should include the names (see figure below). The time information in the left-most column can appear in a variety of formats including "yyyy/m/d", "yyyy/m/d HH:MM", "yyyy/mm/dd", "yyyy/mm/dd HH:MM", or other similar formats. The current version does not support categorical variables; users may need to convert such data to an acceptable format before analysis.

	A	B	C	D	E	F
1	DATE	TKN (mg/L)	NH3-N (mg/L)	BOD5 (mg/L)	CBOD5 (mg/L)	Raw flow (MGD)
2	1/1/2002	25.06	14.23	123.59	79.62	180.44
3	1/2/2002	30.77	16.36	153.17	84.46	182.58
4	1/3/2002	23.67	13.86	111.09	55.31	181.08
5	1/4/2002	22.16	11.58	112.51	59.57	181.97
6	1/5/2002	22.50	11.91	122.35	62.59	183.5
7	1/6/2002	22.33	11.58	119.61	104.92	191.81
8	1/7/2002	23.38	11.79	142.09	84.73	190.34
9	1/8/2002	22.21	13.09	122.24	85.22	191.96
10	1/9/2002	18.84	10.98	105.87	88.02	172.53

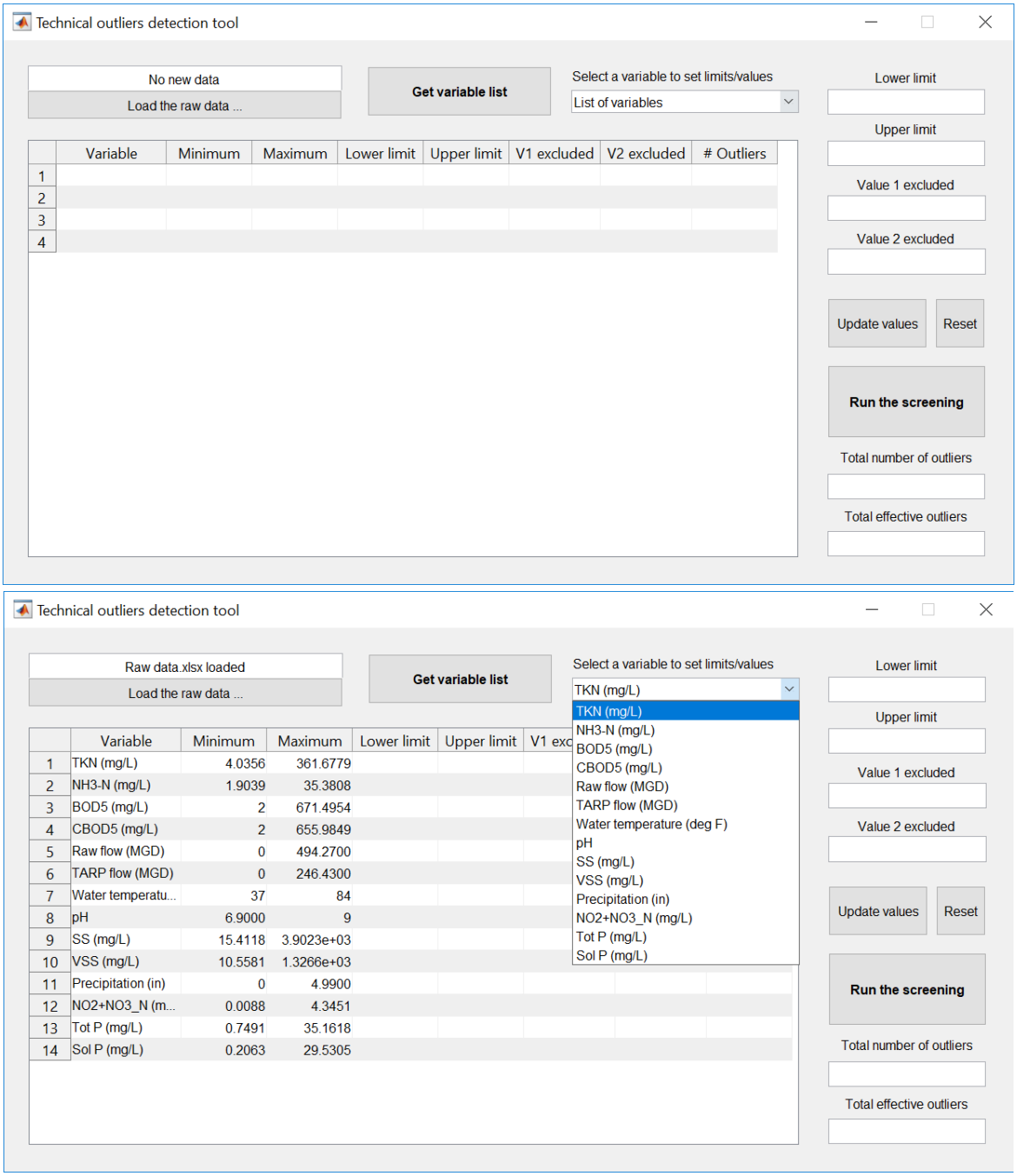
3.2 TECHNICAL OUTLIER ANALYSIS

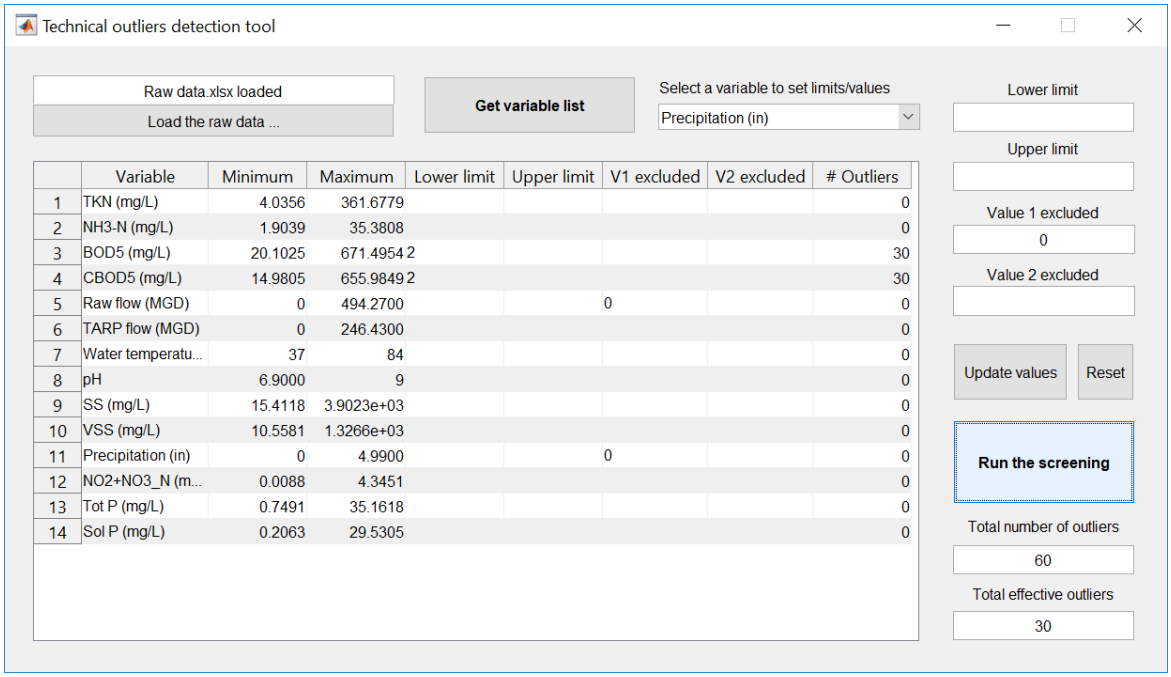
General procedures for this tool are summarized below, followed by figures:

1. Load the raw dataset; get the list of variables and calculate the maximum and minimum values of each variable by clicking the button "Get variable list";
2. Set thresholds (low and upper limits) to identify outliers. For example, if a variable should always be higher than 2, select that variable from the list, enter 2 as the lower limit, and click the button "Update values". Lower and upper limits can be updated at the same time. When define limits to new variables, repeat the similar steps.
3. In addition to remove a range of outliers, users can also exclude specific values form the dataset. Users can define up to two values that need to be excluded from the dataset for each variable at one time. If there are more than two values that need to be excluded, restart the tool and import and screen the new treated dataset for several iterations. If a wrong value is set, it is possible to clean all the values and redefine them by clicking the "Reset" button;
4. Once the above procedures are complete, click the button "Run the screening" to initiate technical outlier detection and exclusion. After running the program, the number of outliers will be shown on the interface. The

last column of the table will show the number of outliers for each variable. The total number of outliers will be shown in the right bottom corner of the interface. With temporal data, it is possible that more than one variable will be identified as outliers for a specific time. The output identified as "total effective outliers" shows the number of observation time that haves at least one outlier variable.

5. When the analysis is complete, five Excel documents will be created. "Technical outlier_Logic values 1.xlsx" and "Technical outlier_Logic values 2.xlsx" are reference excel documents for program computation. "Technical outlier_Summary.xlsx" includes the data exhibited in the tool's table. "Technical outlier_Treated data (for further treatment).xlsx" is the treated dataset and will be used as input dataset for the following dataset pretreatment, such as statistical outlier detection. "Technical outlier_Treated data (for reference).xlsx" is identical to the raw dataset and is used for computation.





	Variable	Minimum	Maximum	Lower limit	Upper limit	V1 excluded	V2 excluded	# Outliers
1	TKN (mg/L)	4.0356	361.6779					0
2	NH3-N (mg/L)	1.9039	35.3808					0
3	BOD5 (mg/L)	20.1025	671.4954	2				30
4	CBOD5 (mg/L)	14.9805	655.9849	2				30
5	Raw flow (MGD)	0	494.2700			0		0
6	TARP flow (MGD)	0	246.4300					0
7	Water temperatu...	37	84					0
8	pH	6.9000	9					0
9	SS (mg/L)	15.4118	3.9023e+03					0
10	VSS (mg/L)	10.5581	1.3266e+03					0
11	Precipitation (in)	0	4.9900			0		0
12	NO2+NO3_N (m...	0.0088	4.3451					0
13	Tot P (mg/L)	0.7491	35.1618					0
14	Sol P (mg/L)	0.2063	29.5305					0

3.3 STATISTICAL OUTLIER ANALYSIS

Application of the different statistical outlier detection tools is similar; a detailed description is only provided for the distance-based method and the other methods will be briefly mentioned.

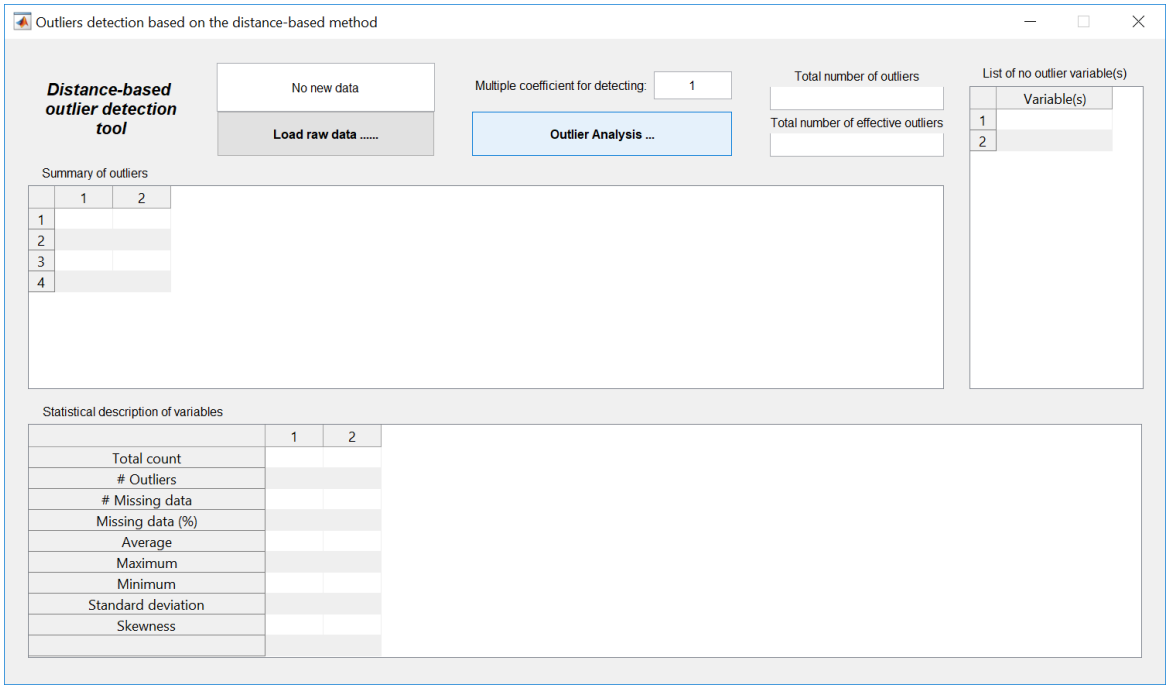
Distance based

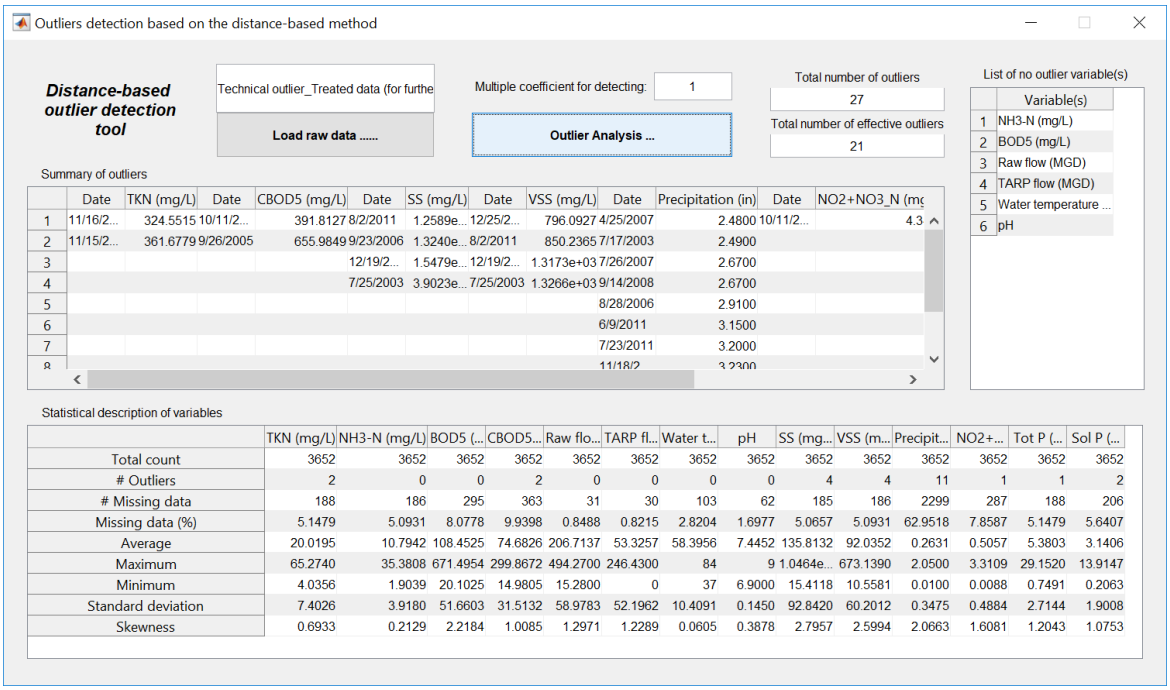
Relative to the other detection methods, the distance-based outlier detection heuristic method typically identifies fewer outliers and it can help to reduce the loss of a large fraction of useful information. The distance-based method can be described as follows (Zhu et al., 2015):

1. Calculate the average value for each variable,
2. Sort the data in an ordered list from the lowest to the highest values, and
3. Calculate the differences between adjacent values in that ordered list.
4. If there is a point in that list where a difference value exceeds the average value of the data, values from that point to the end of the list are considered to be outliers.

Use of this tool:

1. Load the dataset, either as raw data or a dataset that was processed by the technical outlier detection tool (select "Technical outlier_Treated data (for further treatment).xlsx");
2. Click the button "Outlier Analysis…" to obtain the results. It is possible to change the "multiple coefficient for detecting", which is the constant for changing the threshold for outliers. For example, if this coefficient is set equal to 2, outliers will be defined as values that are removed from their nearest neighbor by at least a factor of 2 times the mean value of that variable.
3. There are four types of results appear on the interface. The total number of outliers and the total number of effective outliers are displayed to the right side of the analysis button. Variables without outliers are listed in the table on the far right side. The middle table provides a summary of the outliers, with each variable arranged from lowest to highest value, and the time of each observation is included. The bottom table gives an overall summary on outlier detection and analysis.
4. Five Excel documents will be created. Three of those documents provide summaries of tables from the interface; "Statistical outlier_Treated data.xlsx" is the treated dataset based on the selected statistical detection method; "Outlier_Final treated data.xlsx" is the final treated dataset obtained by removing both technical and statistical outliers and other missing values.

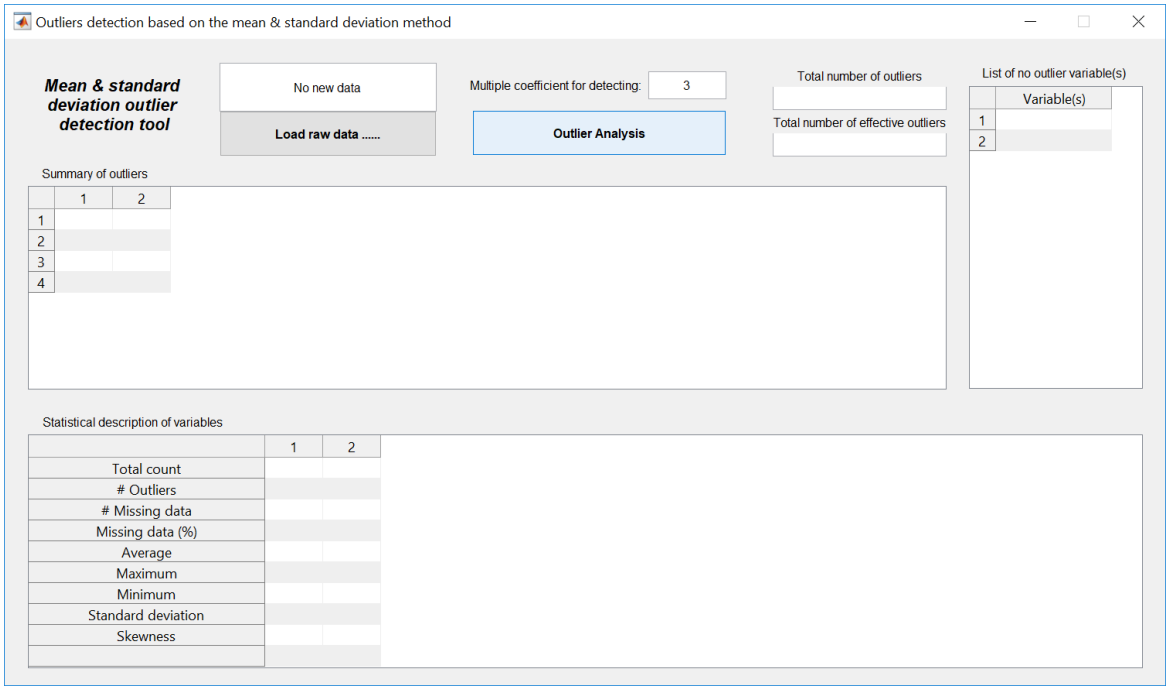




	TKN (mg/L)	NH3-N (mg/L)	BOD5 (...	CBOD5...	Raw flo...	TARP fl...	Water t...	pH	SS (mg...	VSS (m...	Precipit...	NO2+...	Tot P (...	Sol P (...
Total count	3652	3652	3652	3652	3652	3652	3652	3652	3652	3652	3652	3652	3652	3652
# Outliers	2	0	0	2	0	0	0	0	4	4	11	1	1	2
# Missing data	188	186	295	363	31	30	103	62	185	186	2299	287	188	206
Missing data (%)	5.1479	5.0931	8.0778	9.9398	0.8488	0.8215	2.8204	1.6977	5.0657	5.0931	62.9518	7.8587	5.1479	5.6407
Average	20.0195	10.7942	108.4525	74.6826	206.7137	53.3257	58.3956	7.4452	135.8132	92.0352	0.2631	0.5057	5.3803	3.1406
Maximum	65.2740	35.3808	671.4954	299.8672	494.2700	246.4300	84	9	1.0464e...	673.1390	2.0500	3.3109	29.1520	13.9147
Minimum	4.0356	1.9039	20.1025	14.9805	15.2800	0	37	6.9000	15.4118	10.5581	0.0100	0.0088	0.7491	0.2063
Standard deviation	7.4026	3.9180	51.6603	31.5132	58.9783	52.1962	10.4091	0.1450	92.8420	60.2012	0.3475	0.4884	2.7144	1.9008
Skewness	0.6933	0.2129	2.2184	1.0085	1.2971	1.2289	0.0605	0.3878	2.7957	2.5994	2.0663	1.6081	1.2043	1.0753

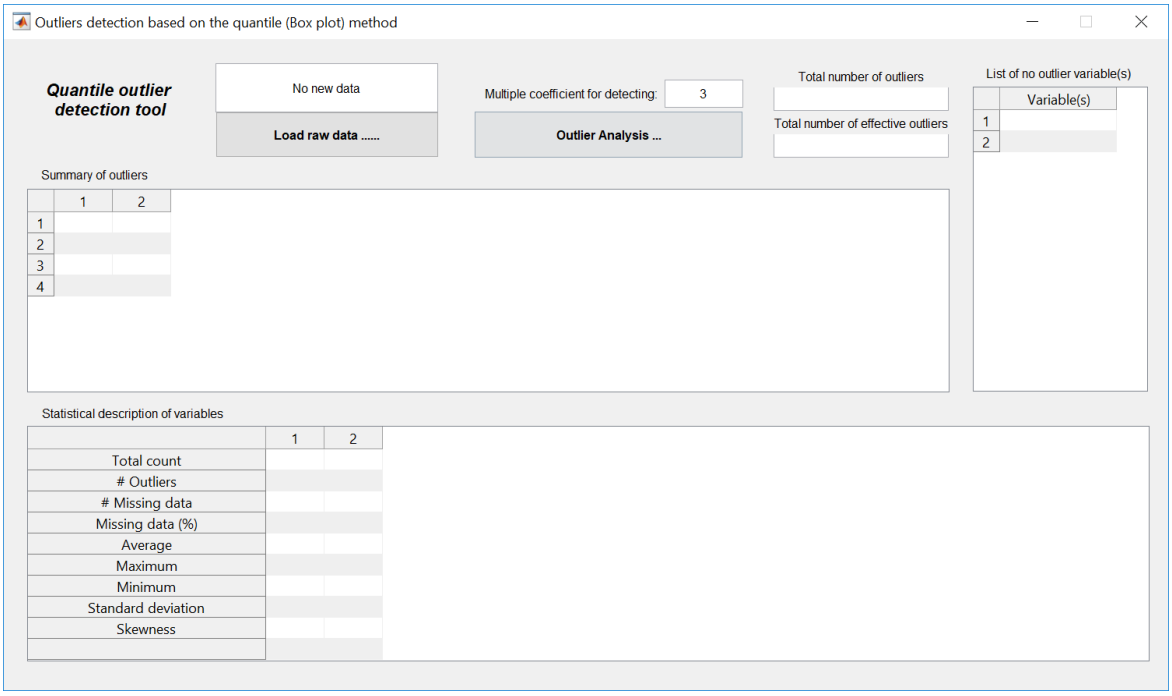
Mean & standard deviation

The default coefficient for detecting outliers based on the mean and standard deviation is 3. In other words, values lower than $\mu - 3\sigma$ or higher than $\mu + 3\sigma$ will be identified as outliers. It is possible to select another value for this coefficient to adjust the number of outliers depending on specific needs.



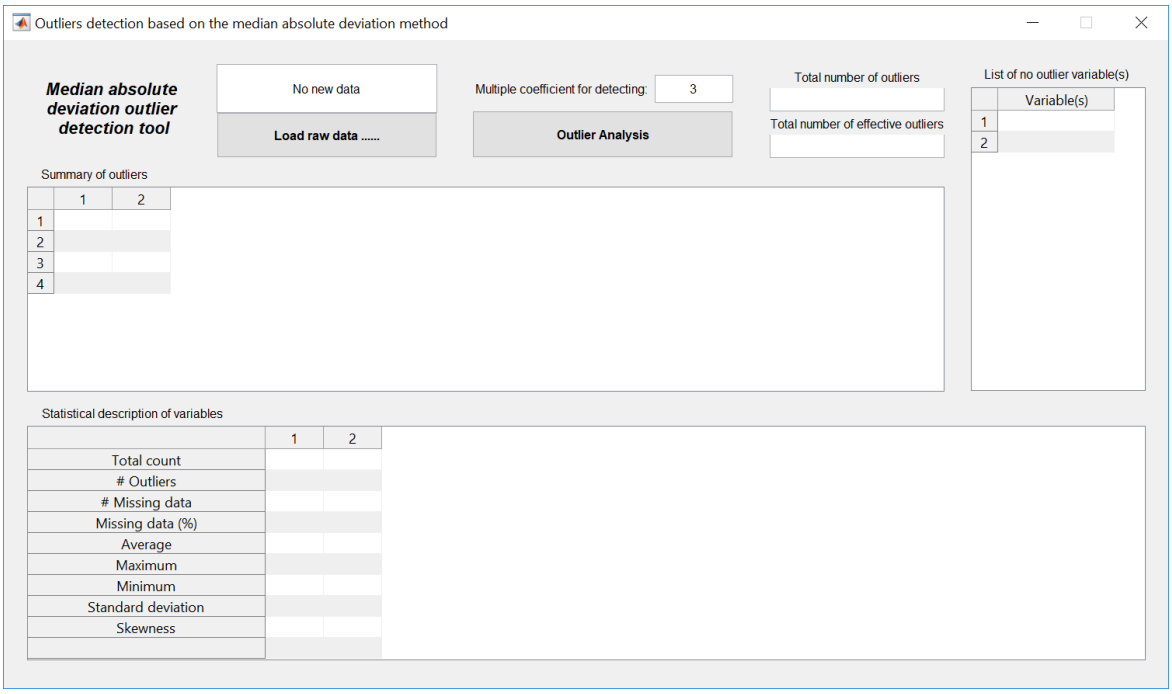
Quantile

Similar to the mean & standard deviation method that the default coefficient for quantile (or box plot) is 3; in other words, any values lower than $Q_1 - 3\times IQR$ or higher than $Q_3 + 3\times IQR$ will be considered as outliers. Users can select a different value appropriate for their specific needs.



Median absolute deviation (MAD)

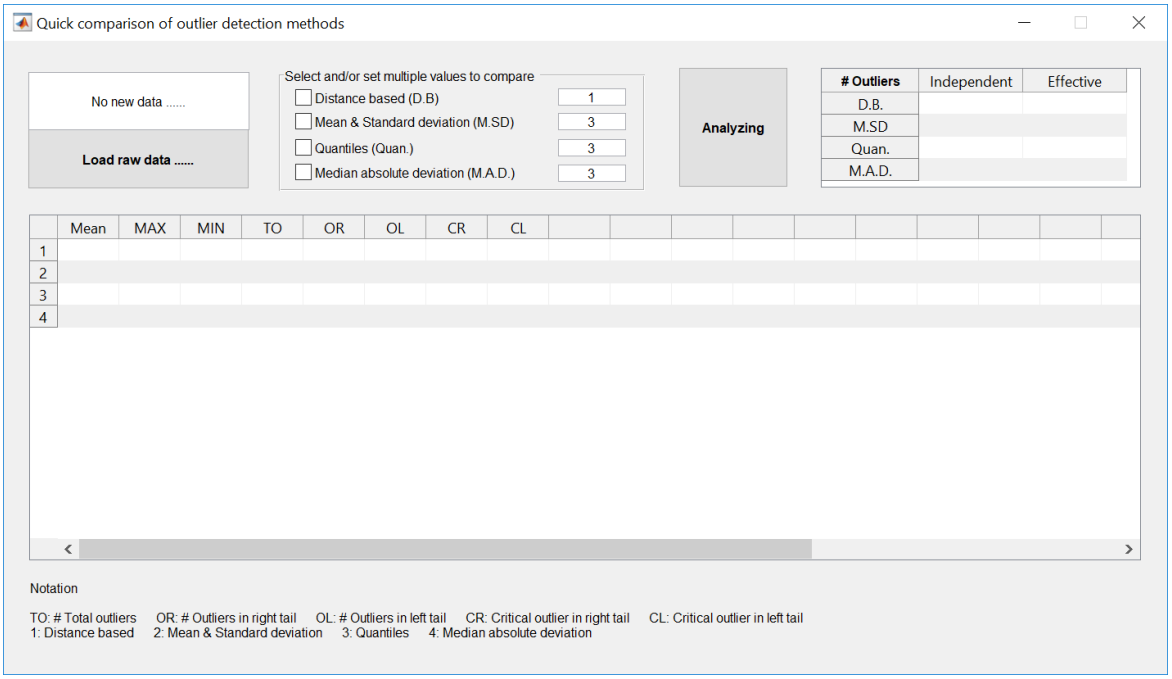
The default coefficient for MAD method is 3, and any values lower than *median* $- 3 \times$ *MAD* or higher than *median* $+ 3 \times$ *MAD* will be considered as outliers. Other values can be assigned to this coefficient.

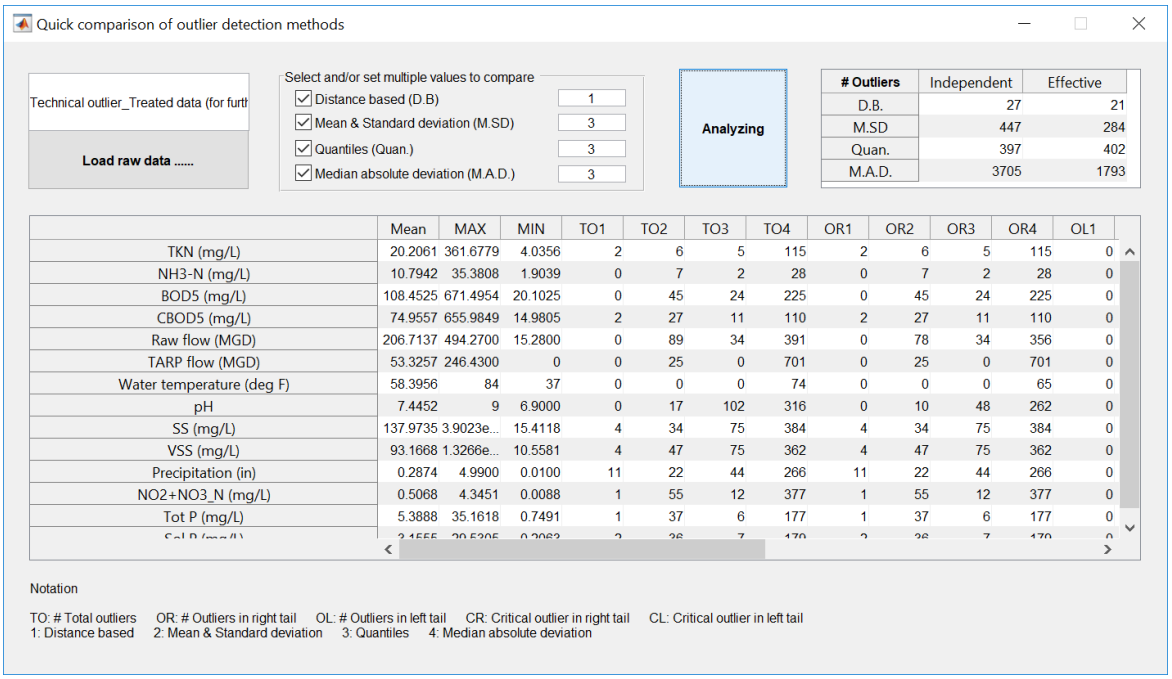


Comparison

To use this tool to provide a preliminary comparison of outlier detection methods:

1. Load the dataset (raw dataset or treated dataset after technical outlier analysis).
2. Select the methods to compare, customize the coefficients if desired, and click the button “Analyzing”.
3. Review the results in the two tables. The left, top table includes the number of independent outliers and the number of effective outliers. The table below summarizes statistical details and outliers for variables.





4. PERIODOGRAM ANALYSIS

The periodogram analysis tool can be used to look for patterns in the time series values for a variable. It can also help to determine a time span window(s) for selecting initial regressors. The three criteria used to decide whether or not a variable has a significant periodic pattern are value-to-peak ratio (VPR), upper limit on the number of VPR, and minimum signal-to-noise ratio (SNR). VPR is the ratio of any value to the peak value of the investigated time period (default = 14 time units, e.g. days) and it defines the threshold that you believe makes it a signal. The default VPR threshold value is 0.2; in other words, values higher than 20% of the peak value are counted as signals. The number of VPR reflects how many of these kinds of signals are found and an upper limit on the number to help avoid too many noisy signals. The minimum SNR allows to screen noises; the user selects this value to include effective signals and exclude noisy values. The SNR is defined as

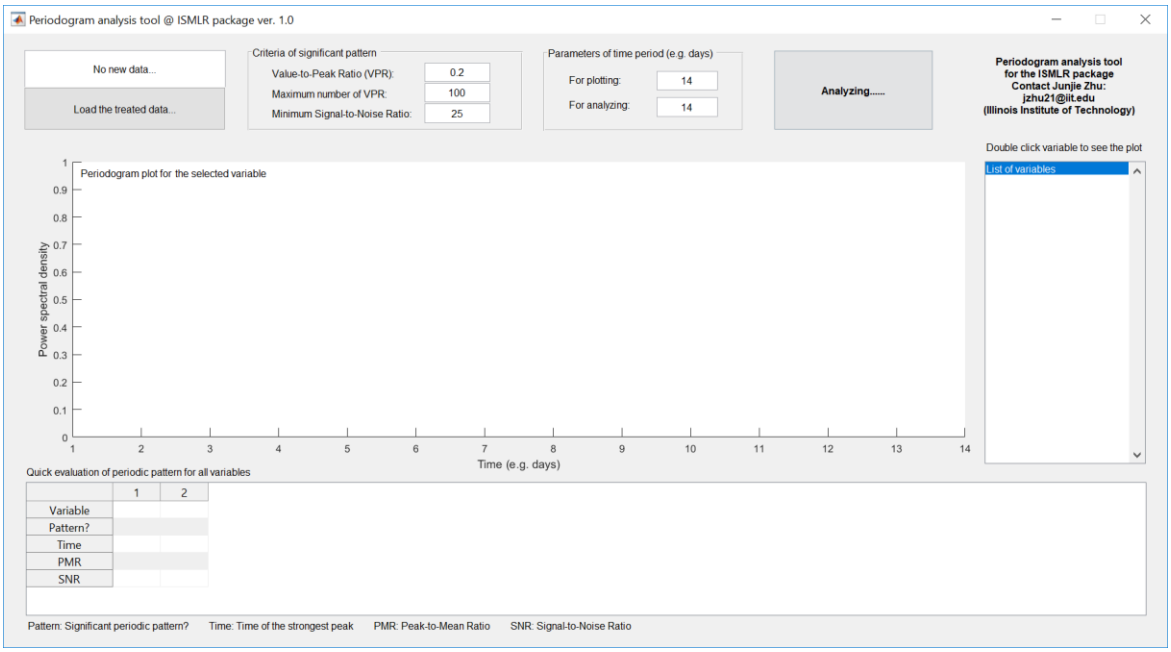
$$SNR = 10 \times \log\left(\frac{Peak\ signal\ value}{Average\ signal\ value}\right)$$

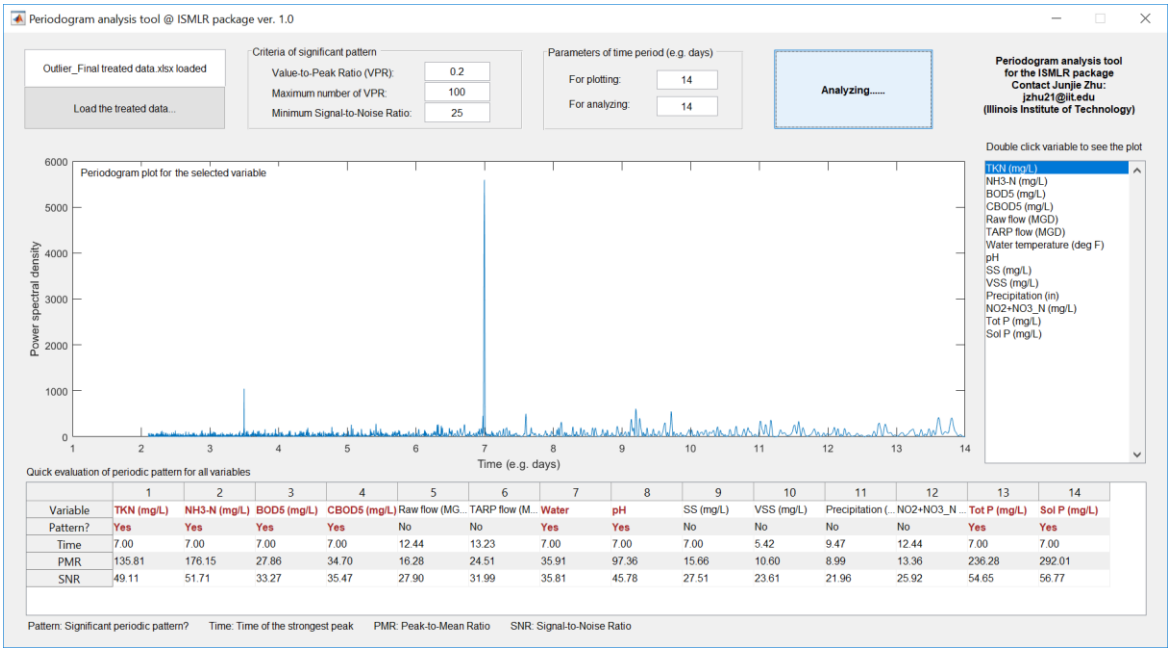
The time span can be specified independently for plotting and analyzing, and their default values are 14.

Use of this tool:

1. Load a dataset (e.g. “Outlier_Final treated data.xlsx”);
2. Use default parameter values or customize them, and click the button “Analyzing……”;
3. The list of variables will be exhibited on the right side; double-clicking any variable provides details of the periodogram plot;
4. The bottom table summarizes the analysis results, including a decision about whether or not there is a significant periodic pattern, the time of the strongest peak, the peak-to-mean ratio (signal value), and the SNR.
5. Two Excel documents will be created. “Periodogram data.xlsx” provides plotting data for all the variables; “Periodogram summary.xlsx” includes the data from the bottom table.

Direction: The decision of pattern and time of strongest peak can help you to determine what will be the best time span window to develop and build time-series regressors.





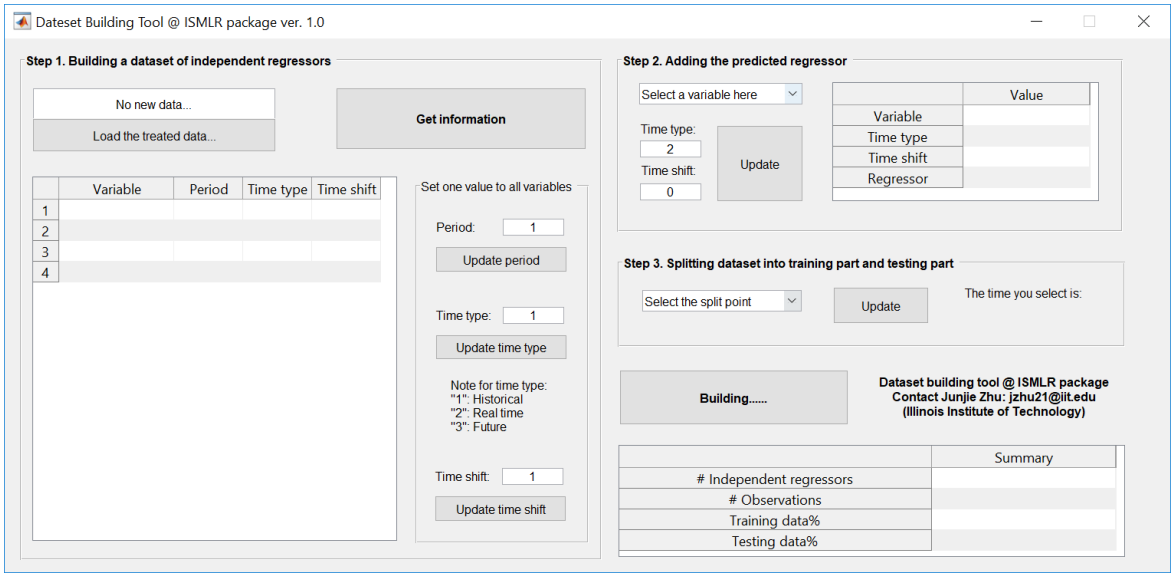
5. DATASET DEVELOPMENT

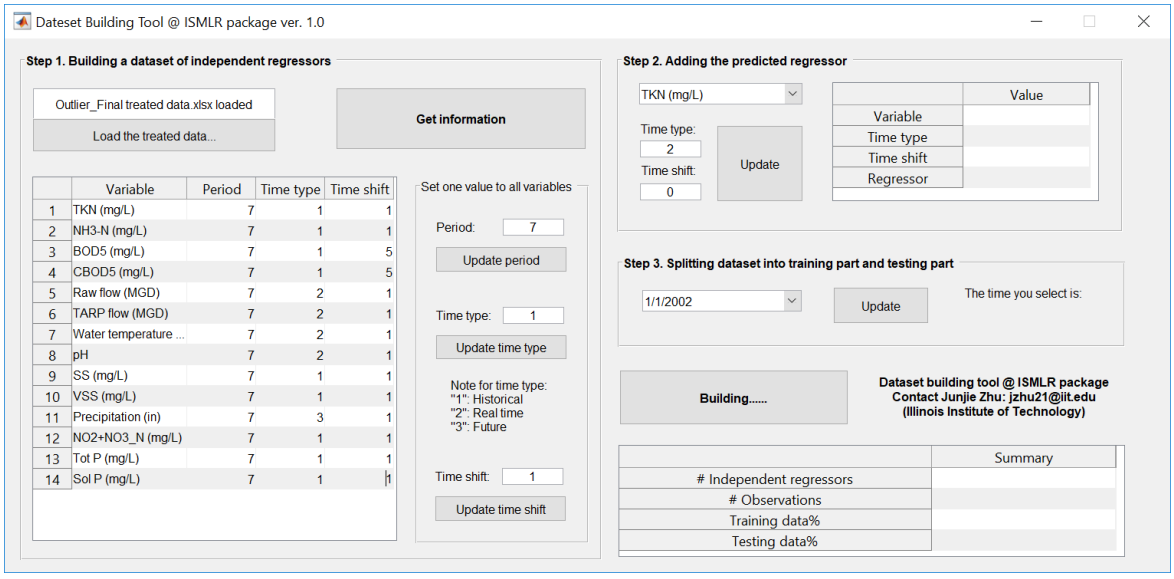
A typical time-series dataset that used for regression analysis includes the same timeframe variables (such as the same day). Alternatively, a better analysis could include different timeframes of variables (referred as time-series regressors), which can be developed from the raw dataset. The dataset development tool is used to provide such a function based on results from the periodogram analysis. The following steps outline procedures for this tool:

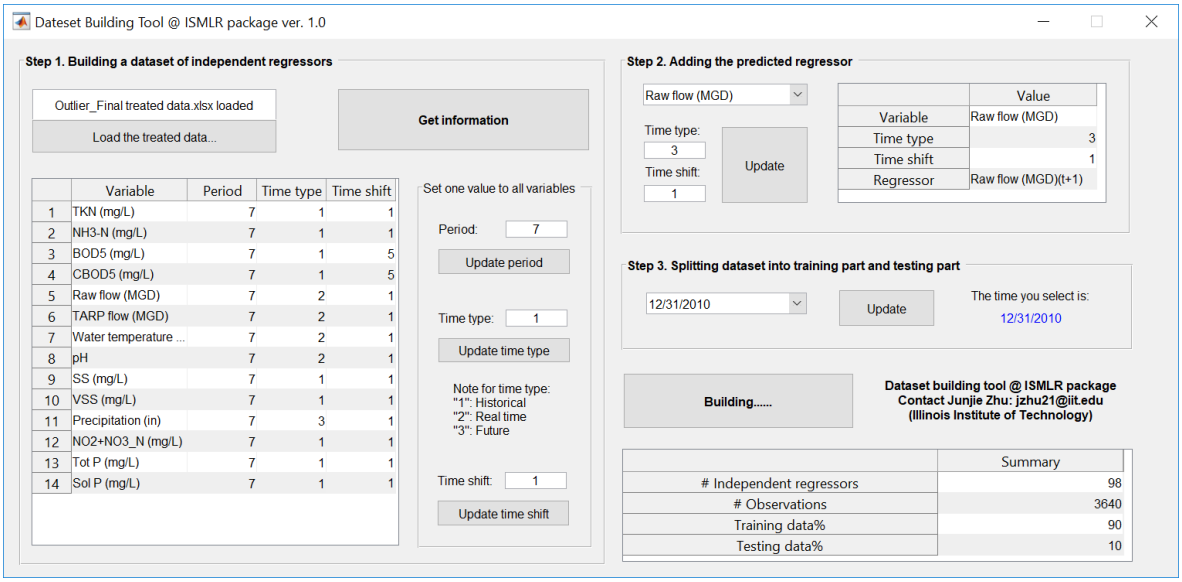
1. Input a dataset (for example, "Outlier_Final treated data.xlsx"). Get the list of variables by clicking "Get the information" button, and the left bottom table will show the variables and the default values for all cells are 1;
2. Define the time-series regressors by changing the values for *period*, *time type*, and *time shift*. Period is the window time span for a variable with a significant periodic pattern (for example, 7 days). Time type shows whether or not the variable can be measured in a relatively long time (historical), in real-time, or even anticipated in future; for example, the type for BOD_5 (five-day biochemical oxygen demand) is "historical"

because it takes five days to measure, whereas flowrate, which can be measured in real-time, is a real-time type. Enter a corresponding number to each variable to specify the number of time-series regressor(s). Time shift is the time difference from the current day. For example, for BOD_5 with a time difference of five days, the value "5" should be entered in the corresponding cell. If the same value is applied to all the variables user can update all the information by filling a value and clicking update button in the right side of the table;

3. The next step is to define the response regressor. Select a variable from the list, define time type and time shift, and update the selection to determine the response regressor. For example, if the next day's raw flow, *raw flow(t+1)*, is the response regressor, then select raw flow as the variable, time type = 3, and time shift = 1. Corresponding information of the reponse regressor will be shown in the table;
4. Split the dataset into a training part and a testing part by selecting a specific time; then click the update button;
5. Build the final datasets by clicking the building button; a brief summary will appear in the bottom table. The example in the figures shows a dataset development from initial 14 variables to 98 time-series regressors.
6. Treated datasets are exported to three Excel documents for all dataset, training part, and testing part, respectively.

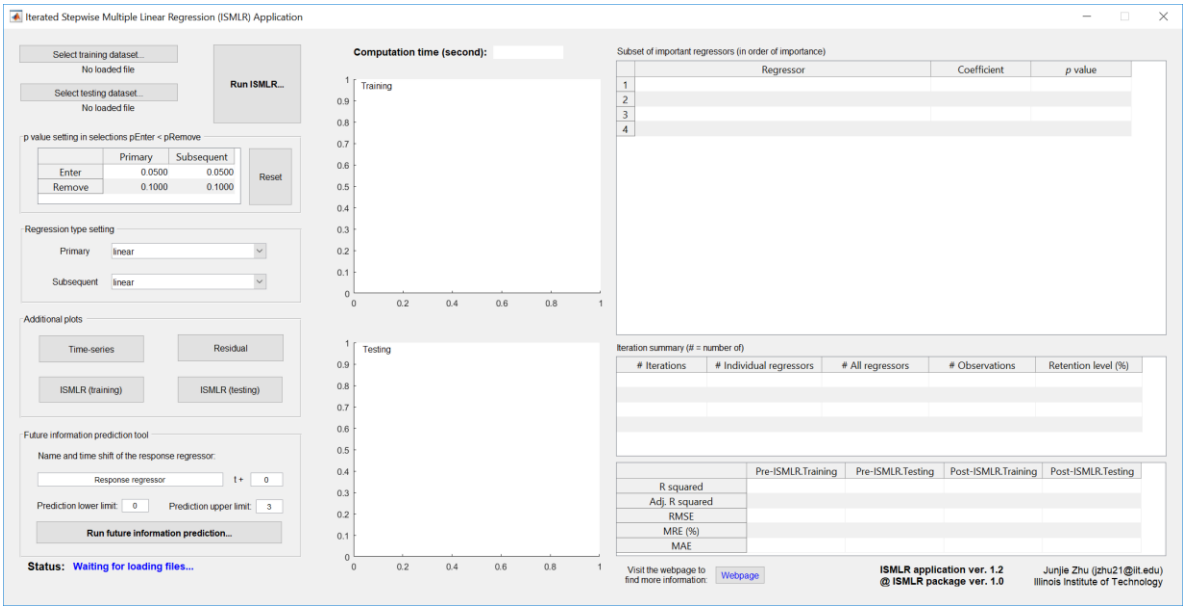






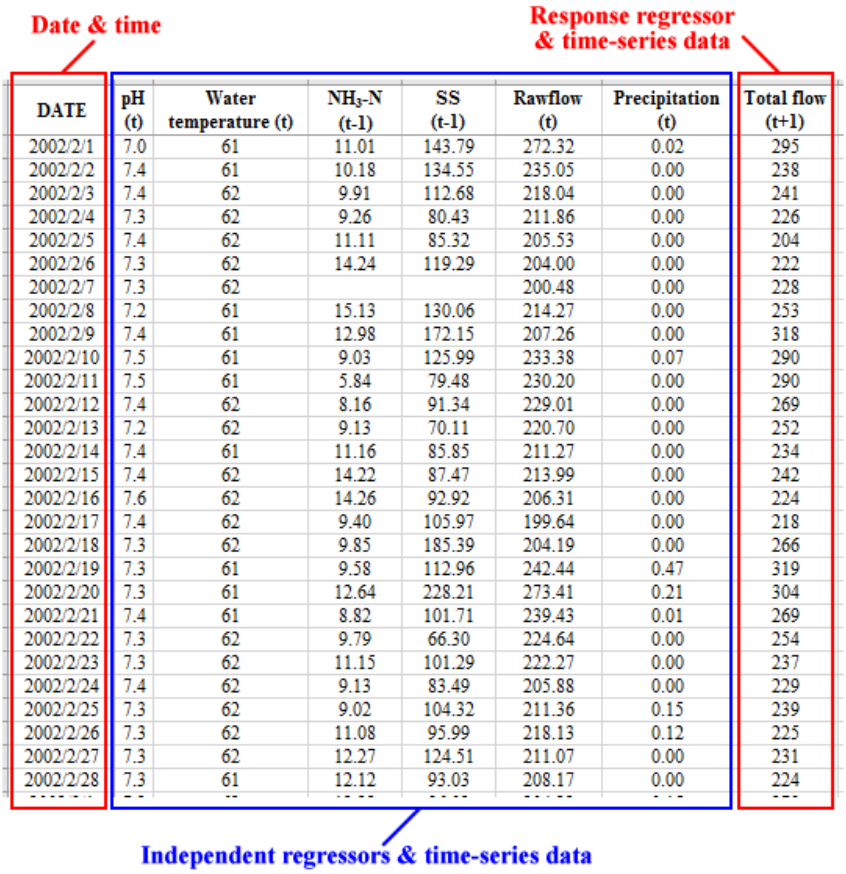
6. ISMLR APPLICATION

The ISMLR application interface is shown below; instructions are described in the following sections.



6.1 INPUT DATASET

The *ISMLR application* is designed to solve time-series regression problems. The input dataset must be prepared in Excel format (*.xls or *.xlsx). Excel documents with treated dataset can be directly used as the input datasets if user uses above built-in data preprocessing tools. The required dataset format includes three elements: Headers, date/time, and data. The date/time information in the left-most column can appear in a variety of formats including “yyyy/m/d”, “yyyy/m/d HH:MM”, “yyyy/mm/dd”, “yyyy/mm/dd HH:MM”, or other similar formats. The response regressor appears in the far-right column. Between date/time and response regressor are the columns of independent regressors. The first row contains headers or regressor names, and the remaining rows are for date/time or data.



DATE	pH (t)	Water temperature (t)	NH_3-N (t-1)	SS (t-1)	Rawflow (t)	Precipitation (t)	Total flow (t+1)
2002/2/1	7.0	61	11.01	143.79	272.32	0.02	295
2002/2/2	7.4	61	10.18	134.55	235.05	0.00	238
2002/2/3	7.4	62	9.91	112.68	218.04	0.00	241
2002/2/4	7.3	62	9.26	80.43	211.86	0.00	226
2002/2/5	7.4	62	11.11	85.32	205.53	0.00	204
2002/2/6	7.3	62	14.24	119.29	204.00	0.00	222
2002/2/7	7.3	62			200.48	0.00	228
2002/2/8	7.2	61	15.13	130.06	214.27	0.00	253
2002/2/9	7.4	61	12.98	172.15	207.26	0.00	318
2002/2/10	7.5	61	9.03	125.99	233.38	0.07	290
2002/2/11	7.5	61	5.84	79.48	230.20	0.00	290
2002/2/12	7.4	62	8.16	91.34	229.01	0.00	269
2002/2/13	7.2	62	9.13	70.11	220.70	0.00	252
2002/2/14	7.4	61	11.16	85.85	211.27	0.00	234
2002/2/15	7.4	62	14.22	87.47	213.99	0.00	242
2002/2/16	7.6	62	14.26	92.92	206.31	0.00	224
2002/2/17	7.4	62	9.40	105.97	199.64	0.00	218
2002/2/18	7.3	62	9.85	185.39	204.19	0.00	266
2002/2/19	7.3	61	9.58	112.96	242.44	0.47	319
2002/2/20	7.3	61	12.64	228.21	273.41	0.21	304
2002/2/21	7.4	61	8.82	101.71	239.43	0.01	269
2002/2/22	7.3	62	9.79	66.30	224.64	0.00	254
2002/2/23	7.3	62	11.15	101.29	222.27	0.00	237
2002/2/24	7.4	62	9.13	83.49	205.88	0.00	229
2002/2/25	7.3	62	9.02	104.32	211.36	0.15	239
2002/2/26	7.3	62	11.08	95.99	218.13	0.12	225
2002/2/27	7.3	62	12.27	124.51	211.07	0.00	231
2002/2/28	7.3	61	12.12	93.03	208.17	0.00	224

The following guides and suggestions for preparing the datasets can help to facilitate a successful prediction:

- The two necessary datasets, training and testing, have to be prepared in a consistent format as described above.
- Categorical values will not work for the ISMLR application. Converting or transforming categorical values to numerical data may solve the problem, but this approach has not been tested in detail.

Examples of training and testing datasets are provided along with the ISMLR application, so users can use these datasets to be familiar with the application or create datasets based on their formats. Data in these datasets were scaled with respect to maximum and minimum (feature scaling).

6.2 SELECTING MODEL PARAMETERS

It is not necessary to specify model parameters, and default settings can be readily adopted. However, these parameters can be specified to pursue a better prediction performance. The first choice is to customize the *p*-value, which quantifies the threshold probability associated with each variable in the regression model. Each of the selection steps is evaluated based on the change of *F*-ratios; the selection criteria, F_{enter} and F_{remove}, are defined using *p*-values for "enter" and "remove", respectively, which describe when a variable should be added or subtracted from a model. Default *p*-values are 0.05 and 0.10 for adding and subtracting, respectively, variables from the regression equation. Theoretically a *p*-value can be any value in the range from 0 to 1.0. In practice, assigning larger *p*-values may require a longer computation time but does not guarantee a better prediction performance (Zhu and Anderson, 2017). The ISMLR application allows for independently set *p*-values for the primary regression and subsequent regression(s). In addition, the *p*-value threshold for adding a variable to a model should always be less than a *p*-value for removing a variable from a model.

The second type of choice is to customize the regression function, which has four options:

- *Linear* (e.g. $y \sim x_1, x_2$),
- *Interactions* ($y \sim x_1, x_2, x_1{:}x_2$),
- *Purequadratic* ($y \sim x_1, x_2, x_1^2, x_2^2$), and

- *Quadratic* ($y \sim x_1, x_2, x_1{:}x_2, x_1^2, x_2^2$).

Relative to the other choices, *interactions* and *quadratic* usually take much more time to compute. As a result, unless there is a significant improvement in the prediction performance, these two regression functions should be used with caution. Similar to the *p*-value, users can set different regression functions in the primary regression and the subsequent regression(s). More detailed information about *p*-values and regression can be found in documentation with MATLAB (2013) or in references such as the work by Montgomery and Runger (2010).

6.3 COMPUTATION TIME

The overall computation time will appear immediately after all computations are completed. Computation time could vary significantly depending on:

- Computer performance (CPU, RAM, hard drive, GPU)
- Computer resource usage (number and type of programs that are running simultaneously)
- Dataset size (number of observations and number of regressors), fraction of missing data, and fractions of training and testing data
- Training option selection, including *p*-values and regression types

Table 1 shows different computation times for predicting the next day's total flowrate, based on different combinations of respective regression types in primary and subsequent regressions (default *p*-values). An example of flowrate prediction using the ISMLR application will be described in a later section. Briefly, ten years of data (2002-2011) were divided into a training part (2002-2010) and a testing part (2011). The raw dataset includes 3652 observations, 105 independent regressors, and one response regressor. The computation time varies from less than one minute to more than 30 minutes. Because this example has a large number of regressors in the raw dataset, the regression types (such as *interactions* and *quadratic*) in the primary regression, can significantly increase the computation time. Because primary regression is mainly used to shrink the big cluster of regressors, it is usually a good idea to choose *linear* or *purequadratic* for that first step.

Table 1. Example of computation times (min) for flowrate prediction. Computation times can vary substantially depending on the regression type in primary and subsequent regressions ($j = 3$).

	Subsequent regression(s)			
Primary regression	*Linear*	*Purequadratic*	*Interactions*	*Quadratic*
Linear	0.3	0.4	1.1	2.4
Purequadratic	0.6	0.6	1.9	2.0
Interactions	10.2	10.1	10.5	12.1
Quadratic	29.0	27.4	29.3	30.9

*Test computer specifications:

CPU: Intel® Core™ i5-2520 M (2.50 GHz); RAM: 8.00 GB, 1333MHz DDR3; GPU: NVIDIA NVS 4200M; Hard drive: 500 GB, 5400 rpm; OS: Windows 7, SP 1 (64-bit)

6.4 OUTPUT INFORMATION AND DOCUMENTS

ISMLR prediction results will be automatically presented as figures and tables in the ISMLR application interface, and the results will be exported to individual Excel spreadsheets. The main interface of the ISMLR application presents a figure showing the predicted values of the response regressor as a function of the measured values, including the training dataset and the testing dataset. Additional output information, which can be accessed using buttons on the interface, includes ordinary time-series predictions, residuals, and pre-/post-ISMLR:

- **Time-series**. Measured data of response regressor are plotted as a time-series based on the final model (treated datasets, no missing data), their corresponding predicted values are also shown in the same figure. A figure is plotted for each of the two (training and testing) datasets.
- **Residual**. Residual values (differences between measured and predicted values) are plotted with their corresponding predicted values. Here again there are two figures, one for each dataset.

- **ISMLR (training)**. Similar to the time-series plots, but these are based on training datasets that include the days with missing data. The *Pre-ISMLR* is the model/plot based on an initial raw dataset, whereas the *Post-ISMLR* is the model/plot based on the final dataset that only includes important regressors.
- **ISMLR (testing)**. Similar to the ISMLR (training), but this button shows the results based on the testing datasets.

In addition to the above plots, three tables are presented in the main interface. They summarize important regressors, iterations, and prediction performance:

- **Subset of important regressors**. The final important regressors are summarized in this table. When other types of regression other than *linear* are selected, second order regressors may be found in the list and they are listed as "A × B" or "A × A" where "A" and "B" represent different individual regressors. The regressors and their coefficients are listed in the order of importance based on their *p*-values.
- **Iteration summary**. This table summarizes the number of iterations, the number of individual regressors, the number of all regressors, the number of observations, and the retention level (%). The number of iterations is the number of times that a treated dataset is built; the number of individual regressors accounts for individual independent regressors; the number of all regressors accounts for all linear, interaction, and pure-quadratic regressors; the number of observations accounts for all valid time-series rows (months, days, hours, or minutes) that include both training part and testing part; the retention level expresses the valid number of observations as a percentage of the total number of observations. Note that the number of observations can change during the modeling process because the initial analysis only includes the training part but the final number accounts for both training part and testing part.
- **Prediction performance**. The performance of the pre-ISMLR and post-ISMLR is evaluated based on five criteria: R^2, adjusted R^2, root mean of squared error (RMSE), mean relative error (MRE, %), and mean absolute error (MAE). These parameters are defined as shown in the following equations:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$Adjusted\ R^2 = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1}\right]$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

$$MRE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

All the above results and data are automatically exported to five Excel spreadsheets, readily available for further use and research. The five Excel documents are:

- **Processed training dataset**. This spreadsheet includes the final treated training dataset (final important regressors and cleaned observations) as well as predicted values of the response regressor and their corresponding residual values.
- **Processed testing dataset**. Similar information as for the processed training dataset, but applied to the testing dataset.
- **Evaluation summary**. This spreadsheet includes the prediction performance summary, subset of important regressors, computation time, user-selected modeling options, and iteration summary.

- **Pre-&Post-ISMLR training part**. This spreadsheet includes time-series measured data and predicted values based on the pre-ISMLR model and the post-ISMLR model for the training part.
- **Pre-&Post-ISMLR testing part**. Similar information as for the above training part, but applied to the testing part.

6.5 EXAMPLE 1

Objective: Predict the next day's total flow, Q_t *(t+1)*, at the MWRDGC Calumet WRP

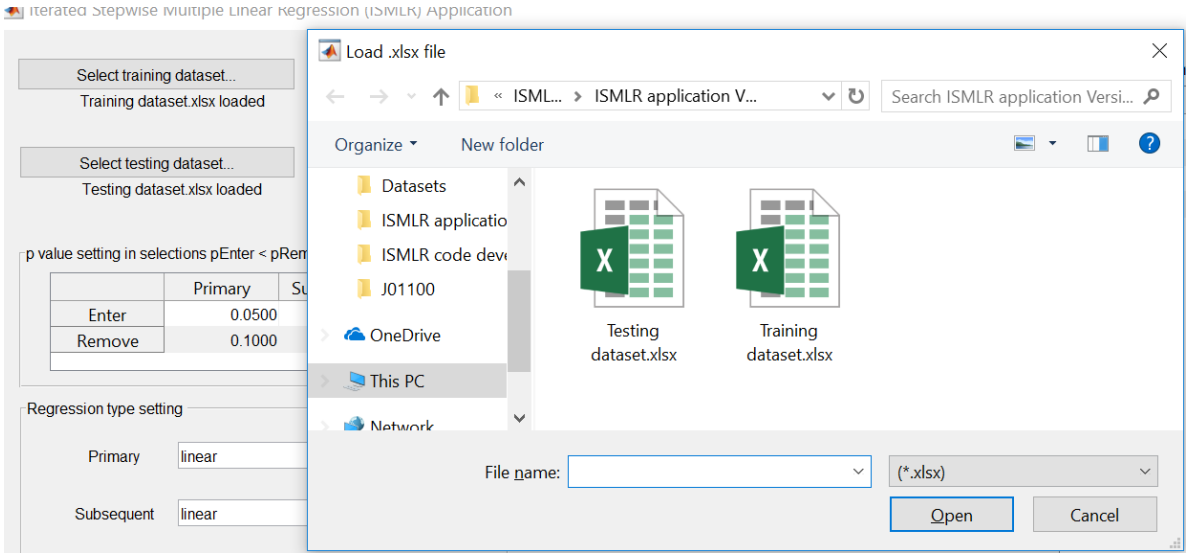
Datasets: Ten years (2002-2011) of historical data at the MWRDGC Calumet WRP, the first nine years of data were the training dataset and the last one year of data was the testing dataset.

Independent variables: 14 variables were used to develop 105 regressors, including 98 historical regressors, five "real-time" (the current day) regressors, and two "future" (the next day) regressors.

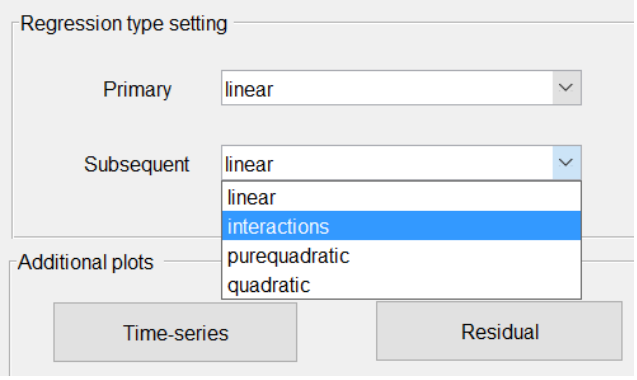
Regression options: Default *p*-values; primary regression type: *Linear*; subsequent regression type: *Interactions*.

Working procedures

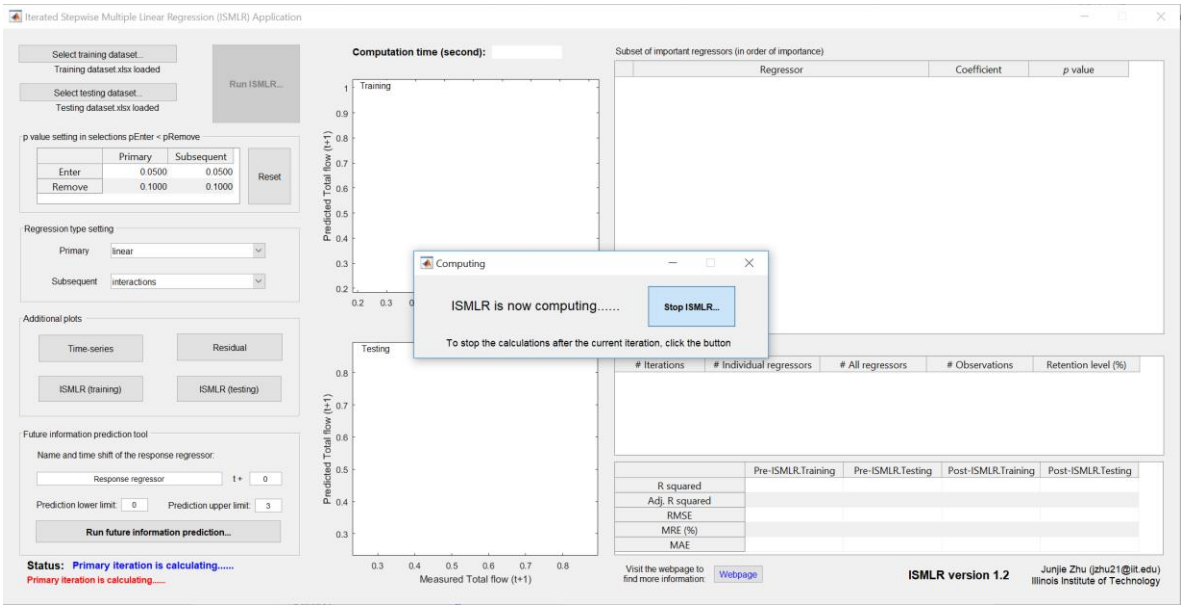
1. Load datasets. The first step is to load the training and testing datasets.



2. Option setting. Choose *interactions* in the subsequent regressions.



3. Run the application. Start the computation by clicking the button "Run ISMLR…".



When the program is running, check the status in the bottom, left corner:

Status: Primary iteration is calculating.....
Primary iteration is calculating.....

Status: Iteration 2 is calculating.....
Primary iteration has been completed (9.386 seconds used)

Status: Iteration 3 or confirmation step is calculating.....
Iteration 2 has been completed (23.955 seconds used)

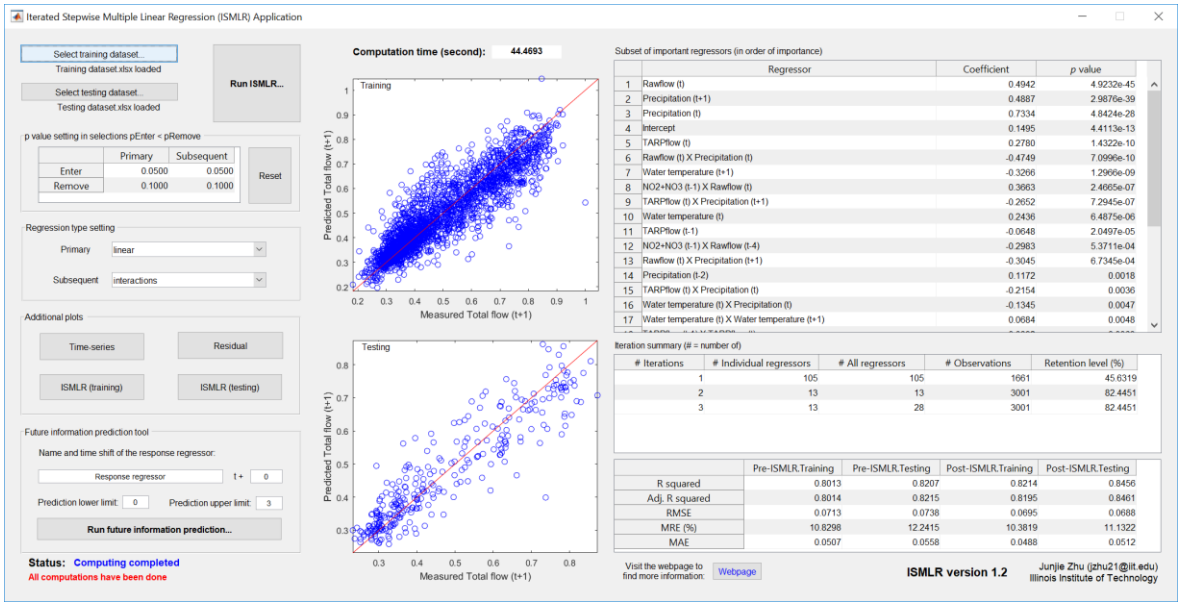
Status: Summarizing, plotting, and exporting data.....
Main function has been completed (38.417 seconds used)

Status: Computing completed
All computations have been done

The iteration summary and prediction performance are updated immediately after each iteration.

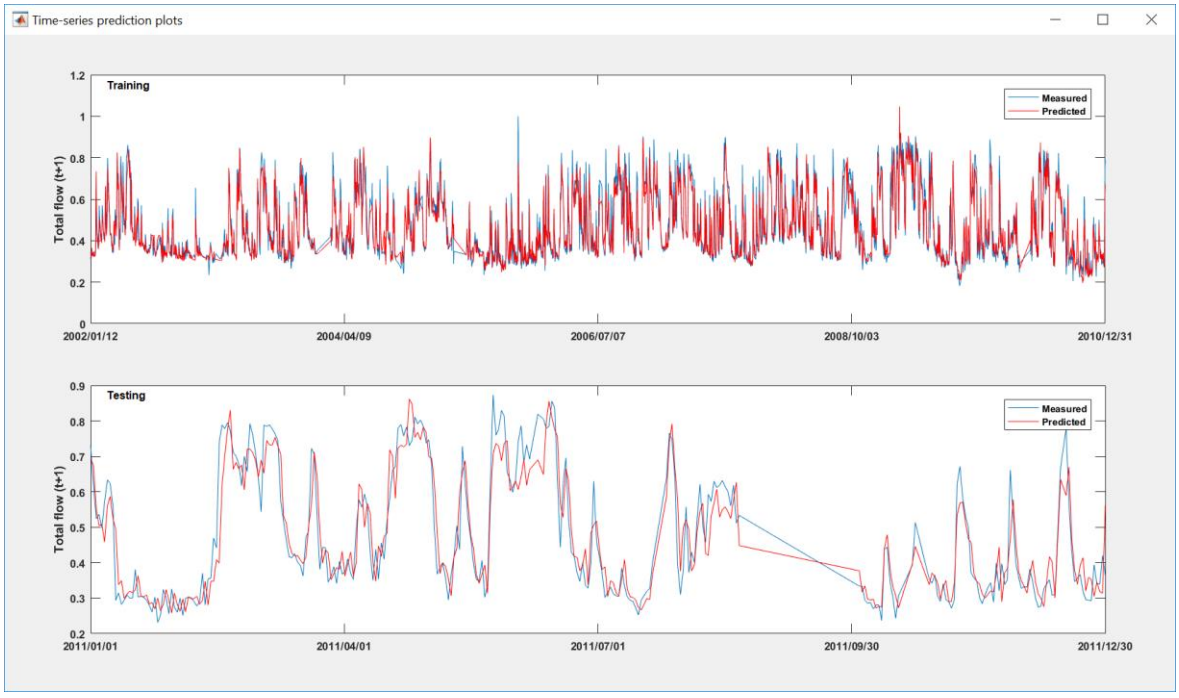
Computation results

1. Main interface

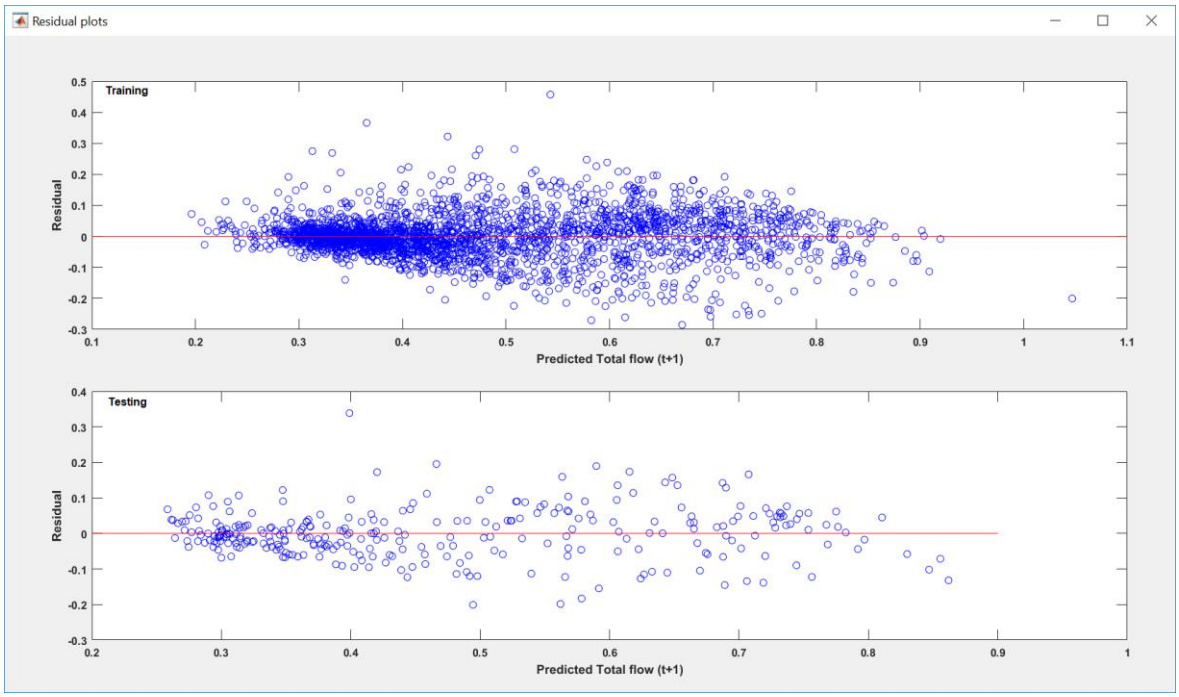


2. Time-series plots

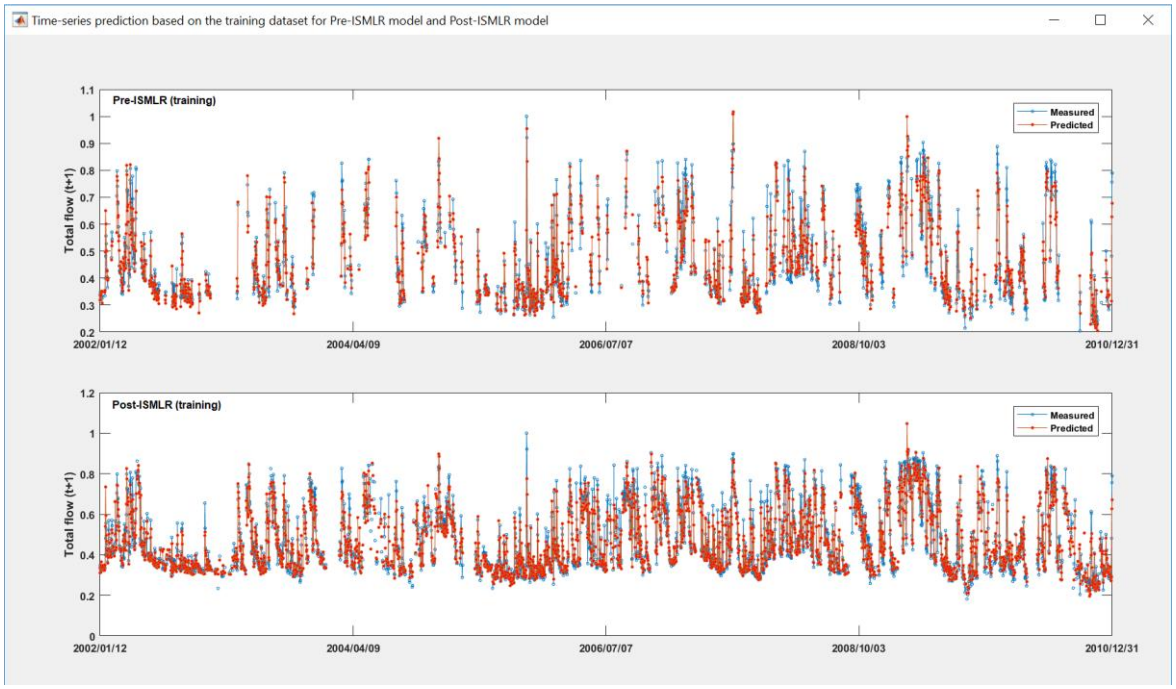
The x-axis of all time-series plots is equally divided into four ranges and given by five corresponding time labels (for example: yyyy/mm/dd).



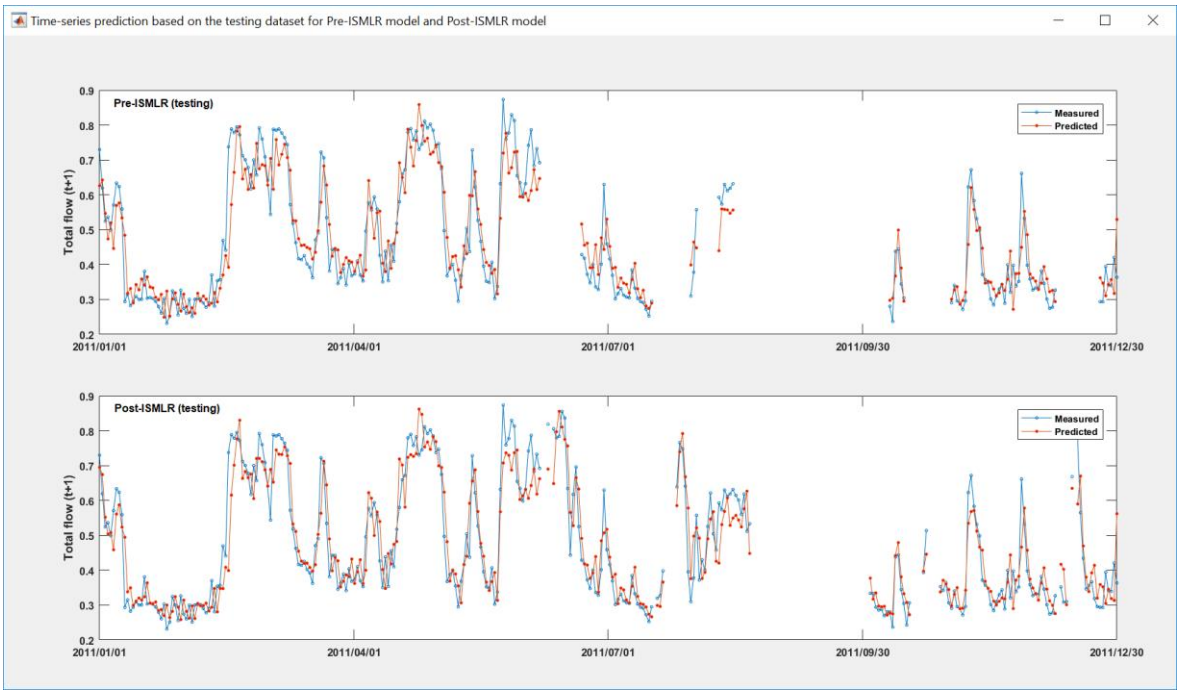
3. Residual plots



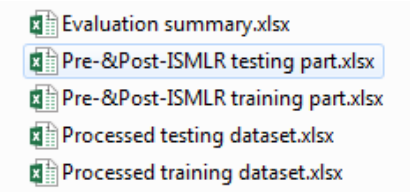
4. Pre-/Post-ISMLR (training) plots



5. Pre-/Post-ISMLR (testing) plots

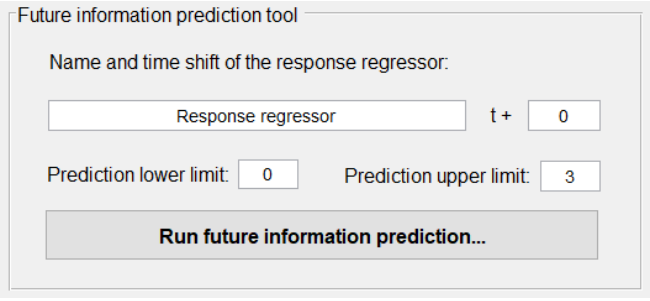


6. Output of Excel spreadsheet documents.



6.6 FUTURE INFORMATION PREDICTION TOOL

A future information prediction (FIP) tool has been included in the ISMLR application since version 1.2. The FIP works similar to the original ISMLR modeling that has options of *p*-values and regression types, and the user-defined modeling options will be applied to predict all future variables. Additional options and inputs are available in the FIP tool:



Two required inputs are the name and time shift of the response regressor. The time shift means the period that the response regressor shifts from the current time. For example, it will be "*t+0*" to predict the current day's influent flowrate (at a WRP) in the original input dataset, whereas it will be "*t+1*" if the response regressor is the next day's influent ammonia concentration.

Two options are prediction lower limit (PLL) and prediction upper limit (PUL), which are used to set the period of the response regressor that the user wants to predict. For example, assume that we are going to predict the next three days' (*t+1 ~ t+3*) influent flowrate while the response regressor in the original dataset is Q_t *(t+1)* (example 1). The PLL and PUL should be set to 0 (t + 1 + 0 = t + 1) and 2 (t + 1 + 2 = t + 3), respectively.

6.7 EXAMPLE 2

Objective: Predict the next seven day's ultraviolet absorbance (UVA), *UVA (t+1) ~ UVA (t+7)*, in the Illinois Fox River

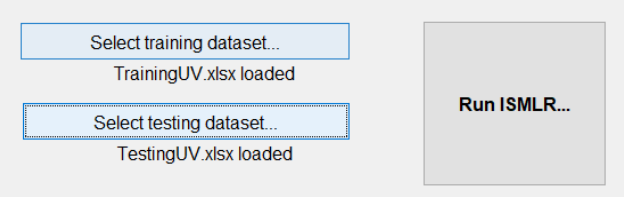
Datasets: Eight years (2002-2009) of historical data from near the City of Elgin, the first six years of data were the training dataset and the last two years of data were the testing dataset. Additional information can be found from Zhu (2012)

Independent variables: Five variables were used to develop 28 regressors, including 22 historical regressors, five "real-time" (the current day) regressors, and one "future" (the next day) regressor.

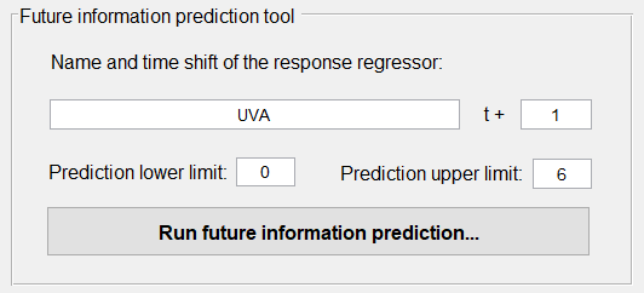
Regression options: Default *p*-values and regression type (*Linear* + *Linear*).

Working procedures

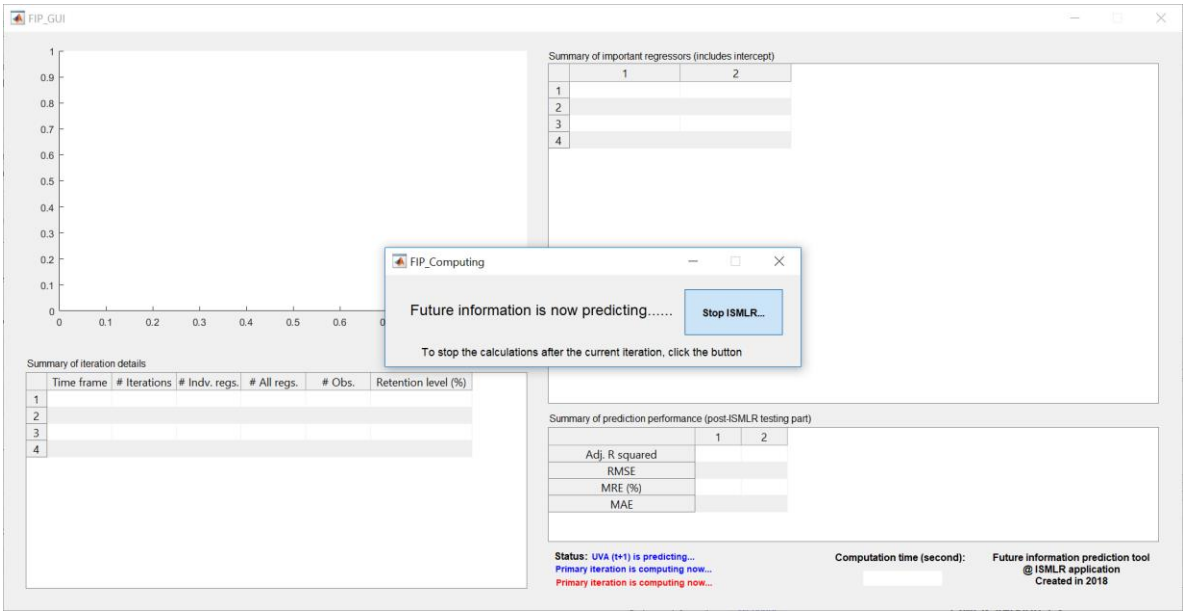
1. Load datasets. The first step is to load the training and testing datasets. In the dataset, the response regressor is *UVA(t+1)*.



2. Option setting and additional information. The existing response regressor in input datasets is UVA (t+1), and PLL and PUL are set to 0 and 6, respectively.



3. Run the FIP tool. Start the computation by clicking the button "Run future information prediction…". A new window will open and prediction results will be updated and shown in the tables iteratively.



When the program is running, check the status at the bottom of the window:

Status: UVA (t+1) is predicting...
Primary iteration is computing now...
Primary iteration is computing now...

Status: UVA (t+1) is predicting...
Iteration 2 is calculating.....
Primary iteration has been completed (1.063 seconds used)

Status: UVA (t+1) is predicting...
Iteration 3 or confirmation step is calculating.....
Iteration 2 has been completed (1.288 seconds used)

Status: UVA (t+1) is predicting...
Summarizing and exporting data.....
Main function has been completed (1.497 seconds used)

.....

Status: UVA (t+4) is predicting...
Primary iteration is computing now...
Primary iteration is computing now...

Status: UVA (t+4) is predicting...
Iteration 2 is calculating.....
Primary iteration has been completed (26.858 seconds used)

Status: UVA (t+4) is predicting...
Summarizing and exporting data.....
Main function has been completed (27.448 seconds used)

.....

Status: UVA (t+7) is predicting...
Primary iteration is computing now...
Primary iteration is computing now...

Status: UVA (t+7) is predicting...
Iteration 2 is calculating.....
Primary iteration has been completed (55.212 seconds used)

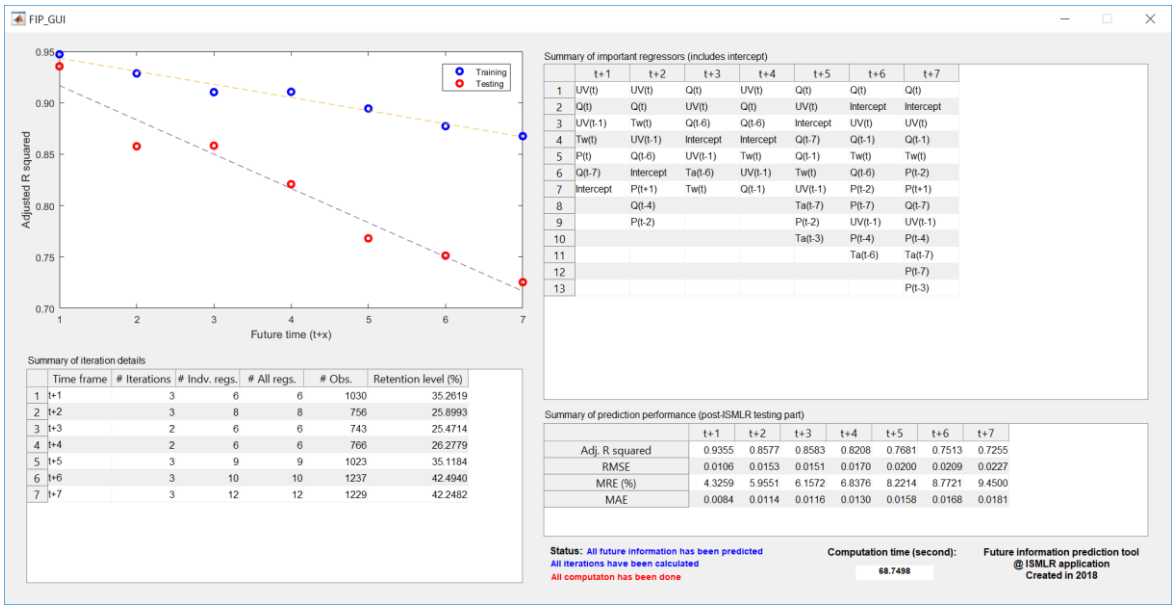
Status: UVA (t+7) is predicting...
Summarizing and exporting data.....
Main function has been completed (58.288 seconds used)

Status: UVA (t+7) is predicting...
This iteration has been done
This iteration has been done

Status: All future information has been predicted
All iterations have been calculated
All computaton has been done

Computation results

1. Main results interface of the FIP tool

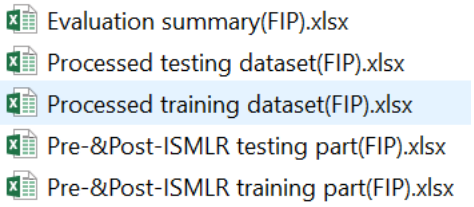


	Time frame	# Iterations	# Indv. regs.	# All regs.	# Obs.	Retention level (%)
1	t+1	3	6	6	1030	35.2619
2	t+2	3	8	8	756	25.8993
3	t+3	2	6	6	743	25.4714
4	t+4	2	6	6	766	26.2779
5	t+5	3	9	9	1023	35.1184
6	t+6	3	10	10	1237	42.4940
7	t+7	3	12	12	1229	42.2482

	t+1	t+2	t+3	t+4	t+5	t+6	t+7
Adj. R squared	0.9355	0.8577	0.8583	0.8208	0.7681	0.7513	0.7255
RMSE	0.0106	0.0153	0.0151	0.0170	0.0200	0.0209	0.0227
MRE (%)	4.3259	5.9551	6.1572	6.8376	8.2214	8.7721	9.4500
MAE	0.0084	0.0114	0.0116	0.0130	0.0158	0.0168	0.0181

In the results interface of the FIP tool, a plot shows the adjusted R^2 values for predicting the response regressor from one day ahead (*t+1*) to seven days ahead (*t+7*) based on the training dataset and the testing dataset. In addition to the plot, three tables provide details of models and specific predictive performance. For example, for *UVA (t+3)*, it has two iterations, the final model includes six individual/all regressors. 743 observations are retained with a retention level of 25.5%. The most important regressor is *Q (t)*. Adj. R^2, RMSE, MRE, and MAE are 0.858, 0.0151 (1/cm), 6.157%, and 0.0116 (1/cm), respectively.

2. Output of Excel spreadsheet documents.



TIPS

The ISMLR application needs to use the resources of MATLAB Compiler Runtime (MCR), so the initial loading of the program may take longer than expected. There could be a delay if:

- A large dataset is being loaded
- Regression type *interactions* or *quadratic* is applied
- System performance is relatively low

The ISMLR application can**not** work if:

- Input training and testing datasets are not consistent
- Row(s) in the dataset(s) include data but not the corresponding date/time

In addition, although it is rare, the ISMLR application cannot work if there are no data missing in the raw datasets and all the regressors are important, because the ISMLR application will enter an infinite SMLR loop. It is possible to work around this problem by manually adding one additional dummy column, giving "1" to all the values and proceed.

Please check all the above items before running the application, so you can avoid these common issues. During computation, you can stop the application by click the "Stop ISMLR" button; the program cannot be stopped during an iteration, but it will stop after the current iteration. Alternatively, you can simply close and restart the ISMLR application. Please send email to the author (jzhu21@iit.edu) if you discover any other problems.

USE AND DISTRIBUTION

Use of the *ISMLR application* for any commercial purpose is prohibited. The ISMLR application is designed for non-profit activities (teaching and research) and it can be downloaded for free for these uses. The author encourages users to spread the program, share their user experiences, and make comments and suggestions. All these steps will improve the ISMLR application.

If you have publications that include results from using the ISMLR application, please include a proper citation. To cite the conventional ISMLR method, use Zhu and Anderson (2016); to cite the peer-viewed paper about ISMLR application, use Zhu and Anderson (2018) (in preparation); to cite the program of ISMLR application or this instruction, use Zhu (2018).

A few final, additional suggestions:

- To download and follow up recent updates of the *ISMLR application*, please visit the author's webpage, Researchgate, or GitHub.
- You can make comments and suggestions using the author's webpage, Researchgate, or via email.
- Discussion via email about potential collaboration is welcome.

Personal webpage: https://junjiezhublog.wordpress.com/

Researchgate: http://www.researchgate.net/profile/Junjie_Zhu4

GitHub: https://github.com/starfriend10/ISMLR-application

Email address: jzhu21@iit.edu

We are also planning to develop ISMLR code using powerful open source software, such as Python or R.

ACKNOWLEDGEMENT

During the development of ISMLR application, I had many discussions with my former advisor, Paul R. Anderson. Paul provided many useful suggestions on the interface designing and language presentation. Robert Nunoo and Boyang Lu are Ph.D. candidates in Anderson's research group; thank for their helps in testing the ISMLR application using their computers.

For the Calumet WRP datasets used in the example 1, I wish to thank Dr. Catherine O'Connor and Judith Moran, Metropolitan Water Reclamation District of Greater Chicago; for providing the data. For the Illinois Fox River datasets used in the example 2, I wish to thank Kyla Jacobsen, Riverside Water Treatment Facility at City of Elgin; for providing the data.

REFERENCES

MATLAB. (2013). Function "*stepwiselm*", introduced in version R2013b, is used to create linear regression model using stepwise regression. https://www.mathworks.com/help/stats/stepwiselm.html (accessed March 22, 2018)

MATLAB. (2017). Document of MATLAB® Compiler™. https://www.mathworks.com/help/compiler/ (March 22, 2018)

Montgomery, D. C., & Runger, G. C., (2010). *Applied statistics and probability for engineers*. Fifth Edition. Published by John Wiley & Sons, Inc. ISBN: 9780471204541

Zhu, J.-J. (2012). *Investigation on the interaction between natural organic matter and calcite*. M.S. Thesis. Illinois Institute of Technology, Chicago, IL. [Official webpage] [Document shared]

Zhu, J.-J., Segovia, J., & Anderson, P. R. (2015). Defining influent scenarios: Application of cluster analysis to a water reclamation plant. *J. Environ. Eng*., DOI: 10.1061/(ASCE)EE.1943-7870.0000934. [Official webpage] [Document shared]

Zhu, J.-J. (2015). *Cyber-physical system for a water reclamation plant: Balancing aeration, energy, and water quality to maintain process resilience*. Ph.D. Dissertation. Illinois Institute of Technology, Chicago, IL. ProQuest/UMI. Publication Number: AAT 3733990; ISBN: 9781339224329. [Official webpage] [Document shared]

Zhu, J.-J. (2018). ISMLR package, version 1.0 was published on July 11, 2018 and can be downloaded from the following sources:
Personal webpage https://junjiezhublog.wordpress.com/,
Researchgate http://www.researchgate.net/profile/Junjie_Zhu4, or
GitHub https://github.com/starfriend10/ISMLR-application (accessed July 23, 2018)

Zhu, J.-J., & Anderson, P. R. (2016). Assessment of a soft sensor approach for determining influent conditions at the MWRDGC Calumet WRP. *J. Environ. Eng*. DOI: 10.1061/(ASCE)EE.1943-7870.0001097. [Official webpage] [Document shared]

Zhu, J.-J., & Anderson, P. R. (2018). Performance evaluation of the ISMLR application for predicting the next day's influent wastewater flowrate at Kirie WRP. In preparation.

Zhu, J.-J., Kang, L., & Anderson, P. R. (2018). Predicting influent biochemical oxygen demand: Balancing energy demand and risk management. *Water Res*., 128(C), 304-313. DOI: 10.1016/j.watres.2017.10.053. [Official webpage] [Document shared]