User manual for the Data Preprocessing Helper Version 0.2.1

Howard Cheung
email: howard.at (at) gmail.com

April 22, 2017

1 Introduction

This tool is made to assist data analysts who are unfamiliar with coding languages to preprocess their time-series data. In a lot of engineering systems, data from different subsystems are obtained differently. While some of them are obtained at different fixed time intervals, some of them, such as on/off signals, are obtained according to time-of-value-change. Some other data contain invalid or missing data points that results in invalid calculation. The resultant data are very difficult for laymen who only use spreadsheet software to analyze their data. This project aims at helping these analysts to preprocess their data by converting the time-of-value-change data to data collected at fixed time intervals.

2 Tutorial

2.1 Conversion of time-of-value-change data to data at fixed time interval

This tutorial gives a quick tour on how to convert a file with time-of-value-change data to a file at fixed time intervals. First we start with a csv file with the structure in Figure 1.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Time	Item 1	Item 2	Item 3	Item 4								
2	1/1/17 7:32:15 AM CST				0								
3	1/1/17 9:33:01 AM CST			1									
4	1/1/17 9:40:10 AM CST			0									
5	1/1/17 10:47:57 AM CST			1									
6	1/1/17 10:56:16 AM CST			0									
7	1/1/17 11:37:51 AM CST			1									
8	1/1/17 11:45:40 AM CST			0									
9	1/1/17 12:24:56 PM CST			1									
10	1/1/17 12:32:31 PM CST			0									
11	1/1/17 1:13:02 PM CST			1									
12	1/1/17 1:21:46 PM CST			0									
13	1/1/17 1:58:57 PM CST			1									
14	1/1/17 2:07:22 PM CST			0									
15	1/1/17 2:27:12 PM CST			1									
16	1/1/17 2:36:06 PM CST			0									
17	1/1/17 3:21:02 PM CST			1									
18	1/1/17 3:29:12 PM CST			0									
19	1/1/17 4:08:07 PM CST			1									
20	1/1/17 4:16:42 PM CST			0									
21	1/1/17 4:58:29 PM CST			1									

Figure 1: Structure of data with time-of-change values

As you can see, there are a lot of unavailable values in Figure 1. The timestamps on the left are also ugly because they are all acquired at different time intervals. It is very difficult to try comparing this set of data side-by-side with the other data set.

To convert the data in Figure 1, we can use the tool provided in this project. The tool is wrapped as an executable so that you don't have to worry about linking your confidential data to the web to do the filtering. Its user interface is shown in Figure 2.

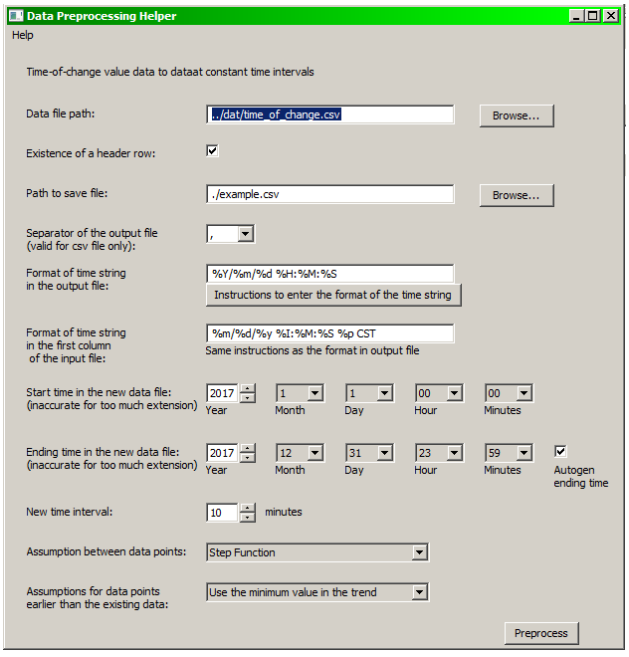


Figure 2: Graphical user interface of the tool

To use the tool, use the first "Browse..." button on the right to choose the data file that you want to convert, and click the second "Browse..." button on the right to choose the directory where you want to save your file. Currently, it supports csv file (Comma-separated Value File), xls file (Microsoft Excel 1997-2003 File) and xlsx file (Microsoft Excel Open XML Format File). The locations of the two "Browse..." buttons are shown in Figure 3.

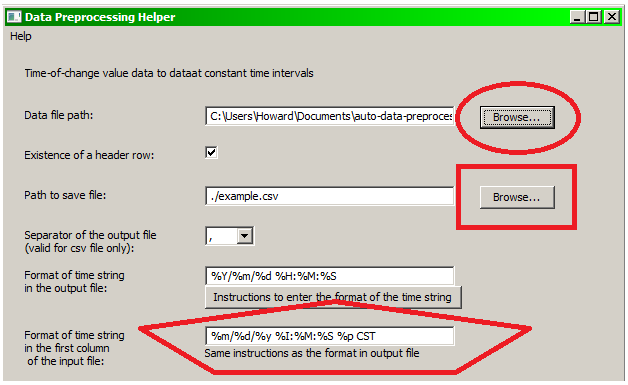


Figure 3: Location of the first "Browse..." button (in a circle), the second "Browse..." button (in a rectangle) and the text box for entering the format of the time string in the data file (in a pentagon)

After that, you need to enter the format of the time string on the leftmost column of your data. The default setting is '%m/%d/%y %I:%M:%S %p CST' supports a format which looks like

12/31/17 7:32:15 AM CST

. Other typical example of time string is shown in Table 1.

Table 1: Sample format string for different types of time string in the data file

Format time string to be entered in the tool	Example time string being support
%m/%d/%y %I:%M:%S %p CST	12/31/17 7:32:15 AM CST 1/15/17 11:33:05 PM CST
%Y/%m/%d %H:%M:%S	2017/12/31 13:00:30 1998/01/32 01:23:02
%y-%b-%d %I:%M %p	99-Jan-01 12:01 AM 00-Feb-31 01:08 PM

Details of their meaning can be found here.

After that, all you need is to press the "Preprocess" button on the right hand corner, and you will have your file when the dialog box in Figure 4 appears.

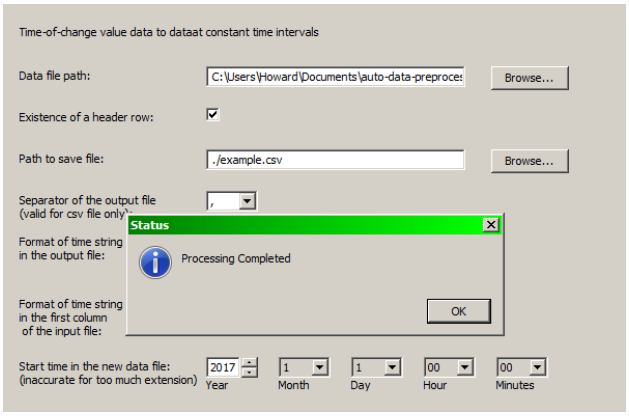


Figure 4: Dialog box showing completion

And you can open the file at your selected location as Figure 5

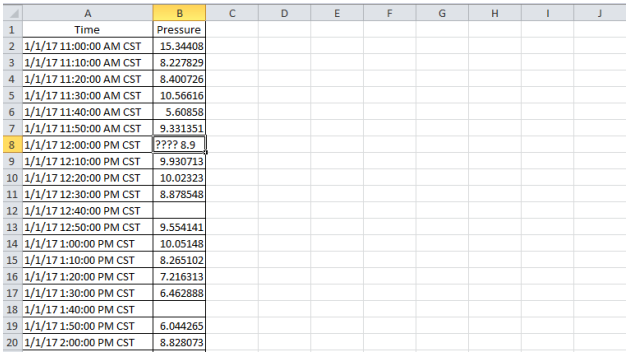
	A	B	C	D	E
1		Item 1	Item 2	Item 3	Item 4
2	1/1/2017 9:00	0	0	0	0
3	1/1/2017 9:05	0	0	0	0
4	1/1/2017 9:10	0	0	0	0
5	1/1/2017 9:15	0	0	0	0
6	1/1/2017 9:20	0	0	0	0
7	1/1/2017 9:25	0	0	0	0
8	1/1/2017 9:30	0	0	0	0
9	1/1/2017 9:35	0	0	1	0
10	1/1/2017 9:40	0	0	1	0
11	1/1/2017 9:45	0	0	0	0
12	1/1/2017 9:50	0	0	0	0
13	1/1/2017 9:55	0	0	0	0
14	1/1/2017 10:00	0	0	0	0
15	1/1/2017 10:05	0	0	0	0
16	1/1/2017 10:10	0	0	0	0
17	1/1/2017 10:15	0	0	0	0
18	1/1/2017 10:20	0	0	0	0
19	1/1/2017 10:25	0	0	0	0

Figure 5: Dialog box showing completion

There are other functions in the tool such as changing the time interval in the output file, using interpolation instead of step function, etc.. While they are available for you to explore in this version of tool, their documentation has not been completed. Details of the functions will be discussed in documentation in future versions.

2.2 Filling in invalid data points in a data file collected at fixed time interval

This tutorial demonstrates how to use the tool to fill in invalid data points collected at fixed time interval. To start the tutorial, we can consider the *missing_data.xls* file as shown in Figure 6.



	A	B	C	D	E	F	G	H	I	J
1	Time	Pressure								
2	1/1/17 11:00:00 AM CST	15.34408								
3	1/1/17 11:10:00 AM CST	8.227829								
4	1/1/17 11:20:00 AM CST	8.400726								
5	1/1/17 11:30:00 AM CST	10.56616								
6	1/1/17 11:40:00 AM CST	5.60858								
7	1/1/17 11:50:00 AM CST	9.331351								
8	1/1/17 12:00:00 PM CST	???? 8.9								
9	1/1/17 12:10:00 PM CST	9.930713								
10	1/1/17 12:20:00 PM CST	10.02323								
11	1/1/17 12:30:00 PM CST	8.878548								
12	1/1/17 12:40:00 PM CST									
13	1/1/17 12:50:00 PM CST	9.554141								
14	1/1/17 1:00:00 PM CST	10.05148								
15	1/1/17 1:10:00 PM CST	8.265102								
16	1/1/17 1:20:00 PM CST	7.216313								
17	1/1/17 1:30:00 PM CST	6.462888								
18	1/1/17 1:40:00 PM CST									
19	1/1/17 1:50:00 PM CST	6.044265								
20	1/1/17 2:00:00 PM CST	8.828073								

Figure 6: Example data file with invalid and missing data points collected at 10-minute interval

Figure 6 shows a file which data contains the following problems

- String characters within a data point at noon
- Multiple data points without any values

To fix the file, we can conduct interpolation with the data adjacent to these problematic data points and fill them in with data from the interpolation. To do so, we open the graphical use interface and choose the followings.

- Data file path: click "Browse..." to choose the file *MissingData.xls*
- Existence of a header row: checked
- Path to save file: click "Browse..." and choose the directory and file name which you want to save the output file
- Format of the time string in the output file: %Y/%m/%d %H:%M:%S (default)
- Format of time string in the first column of the input file: %m/%d/%y %I:%M:%S %p CST
- Start time in the new data file: 2017/01/01 11:00 (the starting time of the file in Figure 6

- Autogen ending time: checked
- New time interval: 10 minutes (same as that of the file)
- Assumption between data points: Continuous variable (inter- and extrapolation)
- Assumptions for data points earlier than existing data: For interpolation, use "Use the first value in the trend"

The output file should have the invalid and missing data points filled while other data points should remain unchanged as shown in Figure 7.

	A	B	C	D	E	F	G	H	I
1		Pressure							
2	1/1/2017 11:00	15.34408							
3	1/1/2017 11:10	8.227829							
4	1/1/2017 11:20	8.400726							
5	1/1/2017 11:30	10.56616							
6	1/1/2017 11:40	5.60858							
7	1/1/2017 11:50	9.331351							
8	1/1/2017 12:00	9.631032							
9	1/1/2017 12:10	9.930713							
10	1/1/2017 12:20	10.02323							
11	1/1/2017 12:30	8.878548							
12	1/1/2017 12:40	9.216345							
13	1/1/2017 12:50	9.554141							
14	1/1/2017 13:00	10.05148							
15	1/1/2017 13:10	8.265102							
16	1/1/2017 13:20	7.216313							
17	1/1/2017 13:30	6.462888							
18	1/1/2017 13:40	6.253577							
19	1/1/2017 13:50	6.044265							
20	1/1/2017 14:00	8.828073							

Figure 7: Resultant data file after interpolation

3 Inquiries

If you encounter bugs about the tool, please send an email to me at howard.at (at) gmail.com or post an issue at the GitHub repository.

4 License

Please refer to the website at the GitHub repository for the most up-to-date information about the licenses.

5 Acknowledgement

The developer(s) would like to acknowledge the followings for the inspiration and resources for the development of the software.

People (in alphabetical order of the family name):

- Dr. Diance Gao at Sun Yat-san University
- Prof. Shengwei Wang at the Hong Kong Polytechnic University
- Mr. KL William Wu at the Hong Kong Polytechnic University

Project (in alphabetical order of the project name):

- Energy Performance Assessment and Optimization on Buildings in PolyU Campus - Stage 1 and Whole Campus at the Hong Kong Polytechnic University