Manual for Setting up the Clustering Algorithm

The pattern clustering algorithm is developed on Python programming language. In its current state, the algorithm is a compilation of Python files. In order to be able to use the clustering algorithm, you will need a development environment and Python installed on your computer. The steps for setting up the environment to use the clustering algorithm are given below;

1. Install Python
   Python is freely available for a large variety of operating systems, including Windows, Macintosh OS and Linux. Required installation files that will make your computer Python-ready can be downloaded from http://www.python.org. The web page also provides instructions on how to install Python. Follow the instructions specified for your operating system.

   Currently, Python has two versions; Python 2.7.5 and Python 3.3.2. The clustering algorithm is developed on Python 2.7.5. Although, the algorithm may work with version 3.3.2 as well, we highly recommend installing version 2.7.5.

2. Install a development environment
   In order to work with the clustering algorithm, you will need an environment that allows you to edit and run Python code. We highly recommend using Eclipse, which is free and open source. The installation files for Eclipse can be downloaded from http://www.eclipse.org/downloads. Eclipse provides a variety of download options, and we recommend the *Eclipse Classic* configuration among these.

   After you install Eclipse, you are recommended to install an add-on that makes Eclipse more Python-friendly. The add-on is called *PyDev*, and the installation instructions can be found at http://pydev.org

3. Installing the necessary Python libraries
   The clustering algorithm utilizes a set of Python libraries such as *numpy*. Instead of installing all the required libraries separately, a much more practical solution is to install the Enthought Python Distribution (EPD). EPD is a bundle of libraries that include all the ones that are required for the clustering algorithm. By installing EPD, you will have all the libraries ready on your computer. EPD can be downloaded from https://www.enthought.com/products/epd/. It is free to download and use EPD with the academic license. The latest installation instructions may also be found on the Enthought website.

4. Importing the clustering algorithm to the Eclipse workspace as a new project
   You will need the zipped file package *BPC.zip,* which is available as an electronic/online supplement to the chapter. After unzipping the archive file, you need to import the folder to Eclipse as a new project.

   Step i. In the File menu of Eclipse, go to the Import command (Figure 0.1).

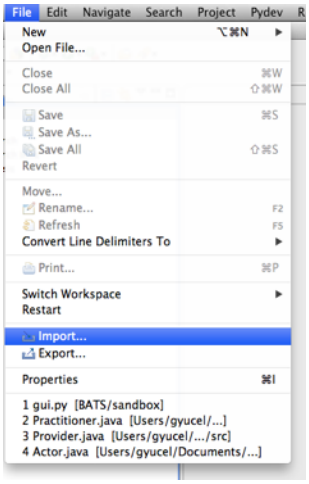

Figure 0.1. Importing BPC to the Eclipse workspace – Step i

Step ii. An import options window will open. In this window, choose the *"Existing Projects into Workspace"* option, which can be found under the General category (Figure 0.2). You do not need to make any changes in the next window. Click *"Finish"* to import the clustering algorithm as a project to your Eclipse workspace (Figure 0.3). Once the importing process is completed, you should be able to see the *"Pattern Clusterer"* as an open project in the navigator pane (Figure 0.4).

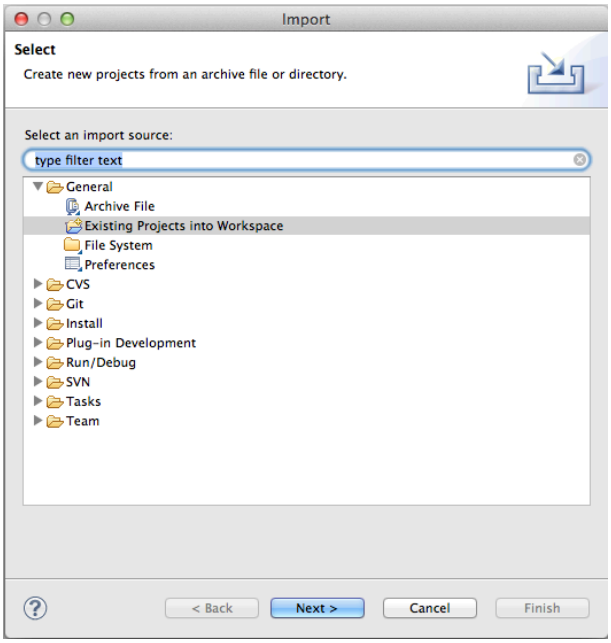

Figure 0.2. Starting the import of BPC to the Eclipse workspace – Step ii

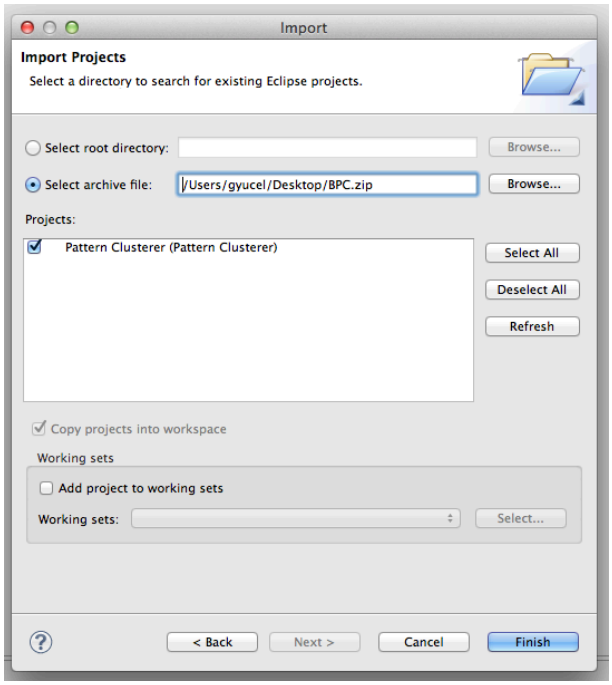

Figure 0.3. Finalizing the import of BPC to the Eclipse workspace – Step ii

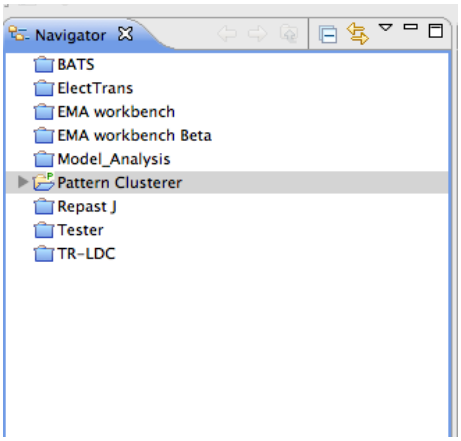

Figure 0.4. Navigator pane after importing the pattern clustering algorithm

Step iii. After importing the algorithm as a project, there is a set of configurations that needs to be done before running the algorithm without compilation errors. On the navigator pane, right click on the *"Pattern Clusterer"* project, and select the "Properties" option from the menu (Figure 0.5). First, you should check if the project path is setup properly. For that, go to "PyDev – PYTHONPATH" and check if "*src"* folder under the *"Pattern Clusterer"* project (i.e. Pattern Clusterer/src) is included in the source folders. If not, add it by using the "Add source folder" option (Figure 0.6).

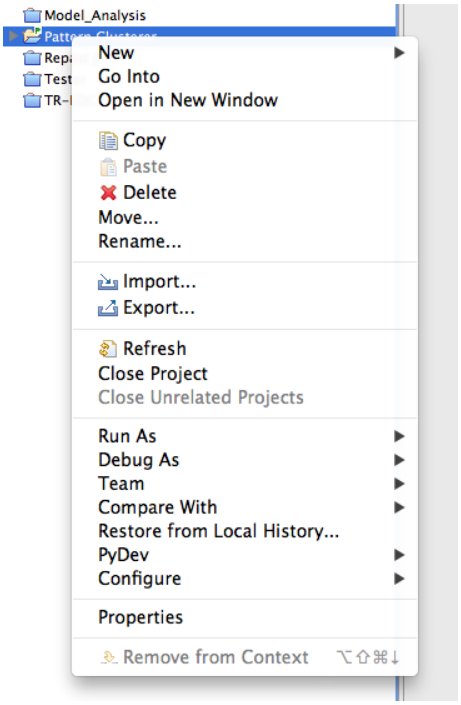

Figure 0.5. Accessing the Properties of the clustering project

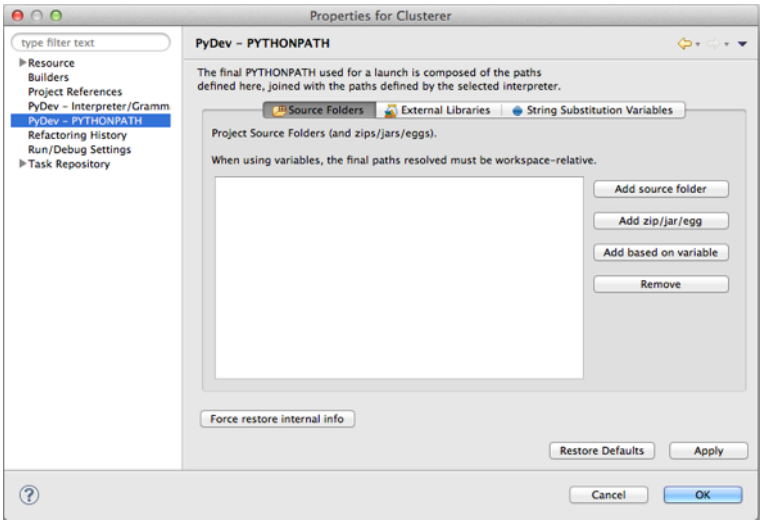

Figure 0.6. Setting up the Source Folders

Second configuration that needs to be done is about "Run Configurations". Go to the "Run" menu of Eclipse, and select the "Run Configurations…" option in this menu. In order to create a new run configuration for the clustering algorithm, click on the "*New Launch Configuration"* (Blank page icon on the top-left corner of the dialog window, marked with an arrow on Figure 0.7).

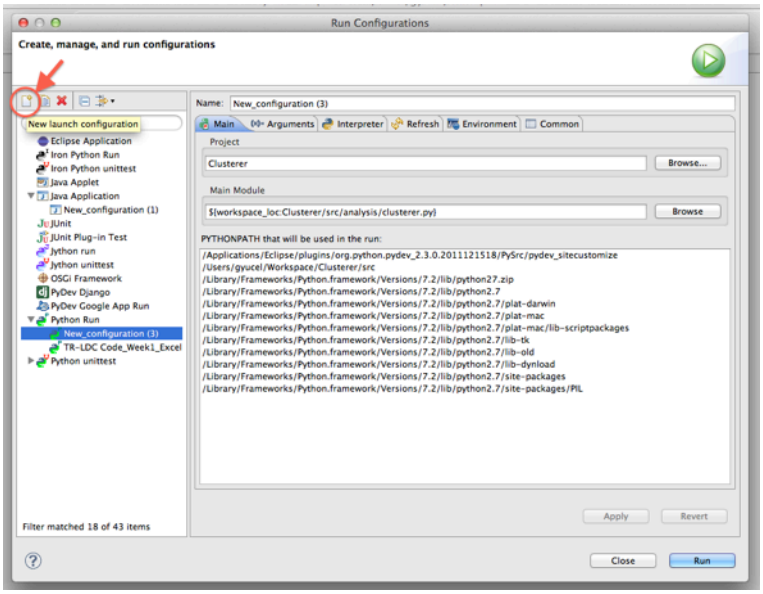

Figure 0.7. Creating a new Run Configuration

You will be asked to provide a name for this new run configuration. After that, select the project using the *"Browse…"* option. You should already have the "*Pattern Clusterer*" project listed in the window that opens when you click the *"Browse…"* button (Figure 0.8). Next, you should specify the *Main Module* of the algorithm. For the algorithm, the main module is named *clusterer.py*, and this file is located in the *analysis* sub-folder of the project. Using the *"Browse"* button, locate and select *clusterer.py* as the *Main Module* of the project (Figure 0.9).

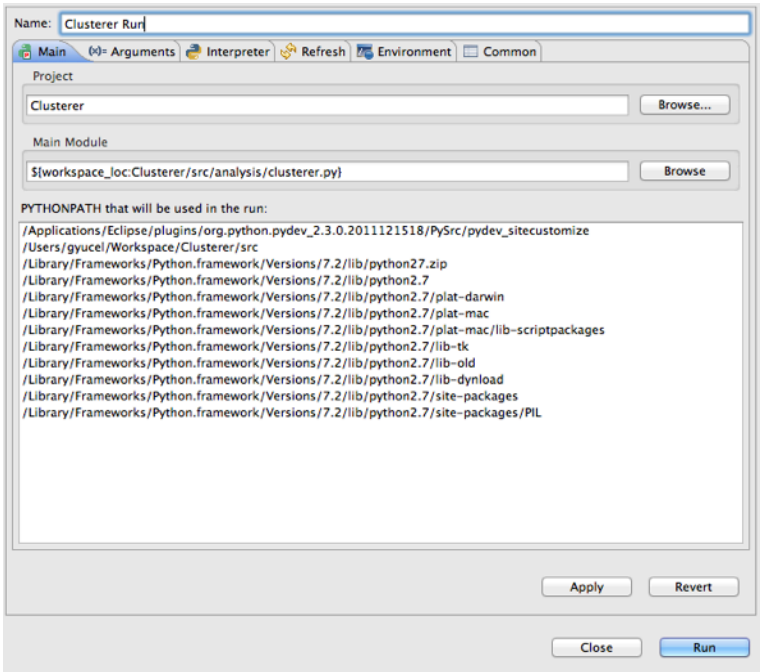

Figure 0.8. Setting up the Run Configuration

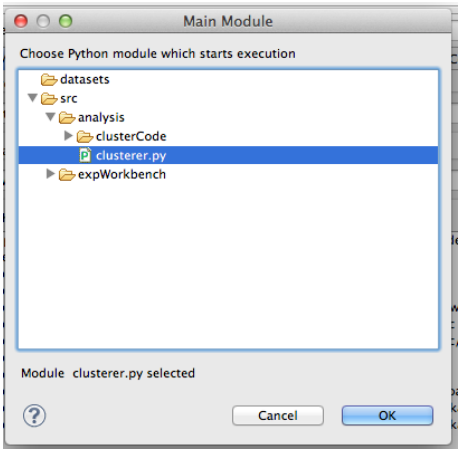

Figure 0.9. Locating the Main Module

In the same window, switch to the "Arguments" tab. You need to change the Working Directory from *Default* to *Other*, and set it to "${workspace_loc:Pattern Clusterer/src/analysis}" (Figure 0.10). Now, you are ready to go. The clustering algorithm should run without any problems.

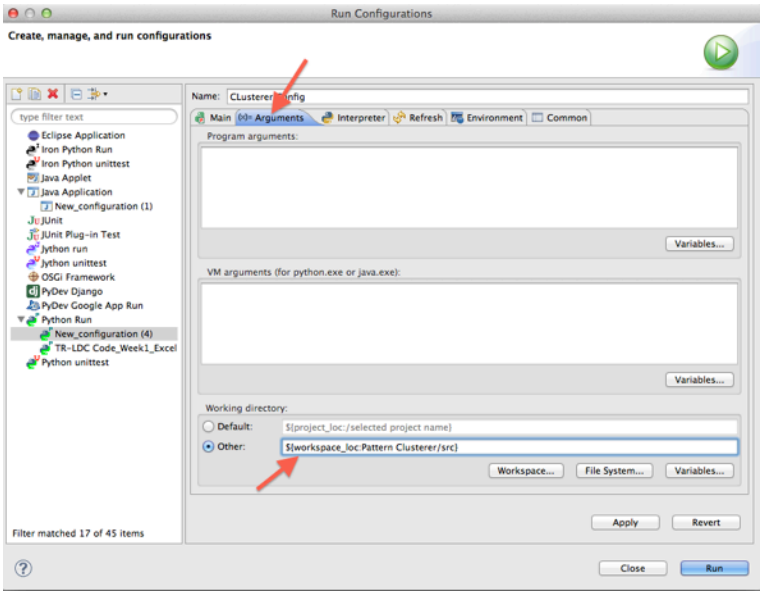

Figure 0.10. Setting the working directory

5. Preparing the model output for analysis

   The algorithm uses a MS Excel (.xlsx) file as the input. Therefore, the model output has to be exported to such a file. There are some details that are crucial for the correct execution of the algorithm. The sheet that contains the data to be clustered should be named "*data*". In the *"data"* sheet, the first row is for headings, and then each row will be treated as a time-series data. It other words, each row corresponds to output from an individual simulation run. The first column is used for the indices of runs. The second, third and fourth columns can be used to store any information about these simulation experiments. The data for the model behavior has to start from the fifth column (column E) as seen in Figure 0.11.

	A	B	C	D	E	F	G	H
1	Simulation	Adjustment I	Initial Temp	Measuremer	T1 Actual Tei	T2 Actual Tei	T3 Actual Tei	T4 Actual Tei
2	1	10	0	5	0	0.5	1	1.49937
3	2	20	0	5	0	0.25	0.5	0.749844
4	3	30	0	5	0	0.166667	0.333333	0.499931
5	4	40	0	5	0	0.125	0.25	0.374961
6	5	50	0	5	0	0.1	0.2	0.299975
7	6	10	10	5	10	10.25	10.5	10.7497
8	7	20	10	5	10	10.125	10.25	10.3749
9	8	30	10	5	10	10.0833	10.1667	10.25
10	9	40	10	5	10	10.0625	10.125	10.1875
11	10	50	10	5	10	10.05	10.1	10.15
12	11	10	20	5	20	20	20	20
13	12	20	20	5	20	20	20	20
14	13	30	20	5	20	20	20	20
15	14	40	20	5	20	20	20	20

data

Figure 0.11. Format of the input file

The Excel file that contains the behaviors to be clustered has to be saved to the "datasets" subfolder in the Eclipse project (Figure 0.12).

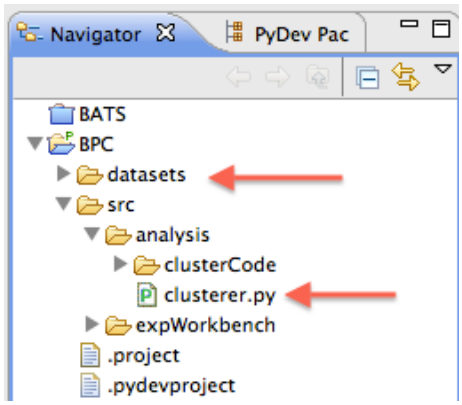

Figure 0.12

6. Running the clustering algorithm

The main component of the algorithm is the clusterer.py file that can be found in the "analysis" sub-folder (Figure 0.12). The only section of this file that is relevant for usage of BPC is the main procedure (the final section that follows the line "if __name__ == '__*main*__':").

Before running the clustering algorithm, the name of the input file has to be specified. In our case, the name of the file is TestModel_Demo.xlsx. Therefore, we set the *inputFileName* variable to *'TestModel_Demo'* (No need to specify the extension).