PANAMA

PRESCRIPTIVE SOLAR ANALYTICS & ADVANCED WORKFORCE MANAGEMENT

D3.1

PV generation profiling methodology

Responsible Partner	University of Western Macedonia
Prepared by	Ioannis Panapakidis Despoina Kothona
Checked by WP Leader	Georgios Christoforidis
Verified by Reviewer #1	Rabia Şeyma Güneş
Verified by Reviewer #2	Onur Enginar
Approved by Project Coordinator	

Project PANAMA is supported under the umbrella of SOLAR-ERA.NET Cofound by the Austrian Research Promotion Agency (FFG), General Secretariat for Research and Technology (GSRT) and the Scientific and Technological Research Council of Turkey (TUBITAK).

Deliverable Record

Planned Submission Date	
Actual Submission Date	
Status and Version	Final version

Version	Date	Author(s)	Notes
Draft version 1	29/12/2020	Ioannis Panapakidis Despoina Kothona	
Draft version 2	25/01/2021	Ioannis Panapakidis Despoina Kothona	
Final version	25/01/2021	Ioannis Panapakidis Despoina Kothona	

Table of Contents

List of Figures

Definition of Acronyms

CI	Calinski-Harabasz Index
CL	Complete Linkage
DBI	Davies-Bouldin Index
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DI	Distortion Index
GUI	Graphical User Interface
MVM	Minimum Variance Method
PV	PhotoVoltaics
SL	Single Linkage
UPGMA	Unweighted Pair Group Method Average
UPGMC	Unweighted Pair Group Method Centroid
WPGMA	Weighted Pair Group Method Average
WPGMC	Weighted Pair Group Method Centroid

Note: Mathematical symbols and terms are explained directly in the corresponding sections.

EXECUTIVE SUMMARY

Deliverable 3.1 titled “PV generation profiling methodology”, includes two major parts: a) The mathematical formulation of the generation profiling methodology and b) the description of the GUI. The first part presents the stages of the methodology. Specifically, it provides information about data representation, clustering algorithms and clustering validation is provided. The methodology includes static and dynamic PV generation profiles. The static profiles are obtained via the application of clustering algorithms. The dynamic profiles are formed utilizing information on weather variables. The second part concerns the GUI description. It serves as a detailed step-by-step description of the functions of the profiling methodology. Thus, the profiling methodology described in the first part is materialized into a user-friendly GUI. The GUI can be distributed as an .exe file and can be run on standard pc systems.

1 INTRODUCTION

The PV generation profiles extraction procedure involves only PV generation data. The generation data are clustered into clusters and the clusters profiles are obtained. Figure 1 presents the input data flow between the Deliverables of the WP3. The data collected from the PV installation are used in all Deliverables. The outputs of the methodology of Deliverable 3.1 are utilized as inputs in Deliverable 3.2 and Deliverable 3.3.

The PV generation profiling methodology consists of several stages, as depicted in Figure 2. The methodology aims at formulating static and dynamic PV generation profiles of a given data set. For the purpose of developing a robust methodology, a variety of clustering algorithms and clustering validity indicators or adequacy measures are considered. This approach allows the user to examine various algorithms of different operation and thus, studying in a more detailed manner a data set and draw more safe and reliable conclusions.

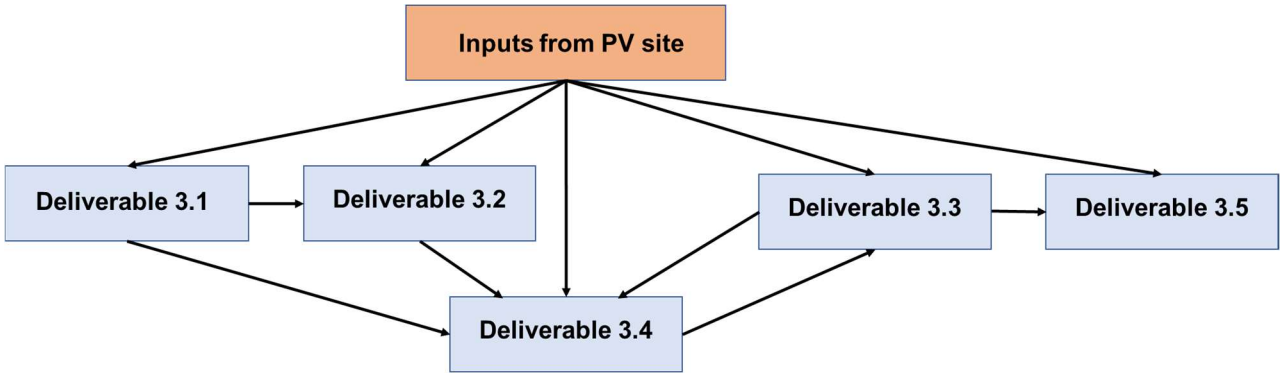


Figure 1. Inputs flow between the Deliverables of WP3.

2 PV GENERATION PROFILING METHODOLOGY

2.1 DATA REPRESENTATION

The first stage of the methodology is data collection and representation. The target variable is the PV generation. We denote with the term "pattern" a daily PV generation curve. Let p_m be the pattern of the m-th day, $m = 1,2,\dots,M$ and M the number of days that the data set refer to. It is expressed as:

$$p_m = [p_m^1, \dots, p_m^t] \quad (1)$$

where $t = 1,2,\dots,T$ is the time step of the metering interval. For instance, if data are collected every 15 min, then $T = 96$. The set that contains the generation data is denoted as $P = \{p_m, m = 1,\dots,M\}$. Clustering involves the similarity of the PV generation curves. Each pattern of the set should be normalized in the proper range before entering in the clustering algorithm. Let p_{max} be the maximum

value of P. The normalization in the [0,1] region is accomplished by dividing each pattern p_m with p_{max}. Let x_m be the normalized pattern. It is:

$$x_m = \frac{p_m}{p_{max}} \quad (2)$$

The set of the normalized patterns is denoted as $X = \{x_m, m = 1, \ldots, M\}$. The scope of the clustering process is to define a set of patterns $C_k = \{c_k, k = 1, \ldots, K, 1 \leq K \leq M\}$ where k is an indicator denoting the k-th cluster and K is the number of clusters. The centroid c_k of the k-th cluster is expressed as the average of all patterns that belong to the cluster:

$$c_k = \frac{1}{M_k} \sum_{m=1}^{M_k} x_m \quad (3)$$

where M_k is the number of patterns that belong to the k-th cluster. Apart from the centroids set, the clustering outcome involves the membership of each pattern to the k clusters. Let u_{km} be the membership of the m-th pattern to the k-th cluster. Let $U = [u_{km}]$ be the partition matrix of set X. The scope of clustering is to group together patterns x_m so that [1]:

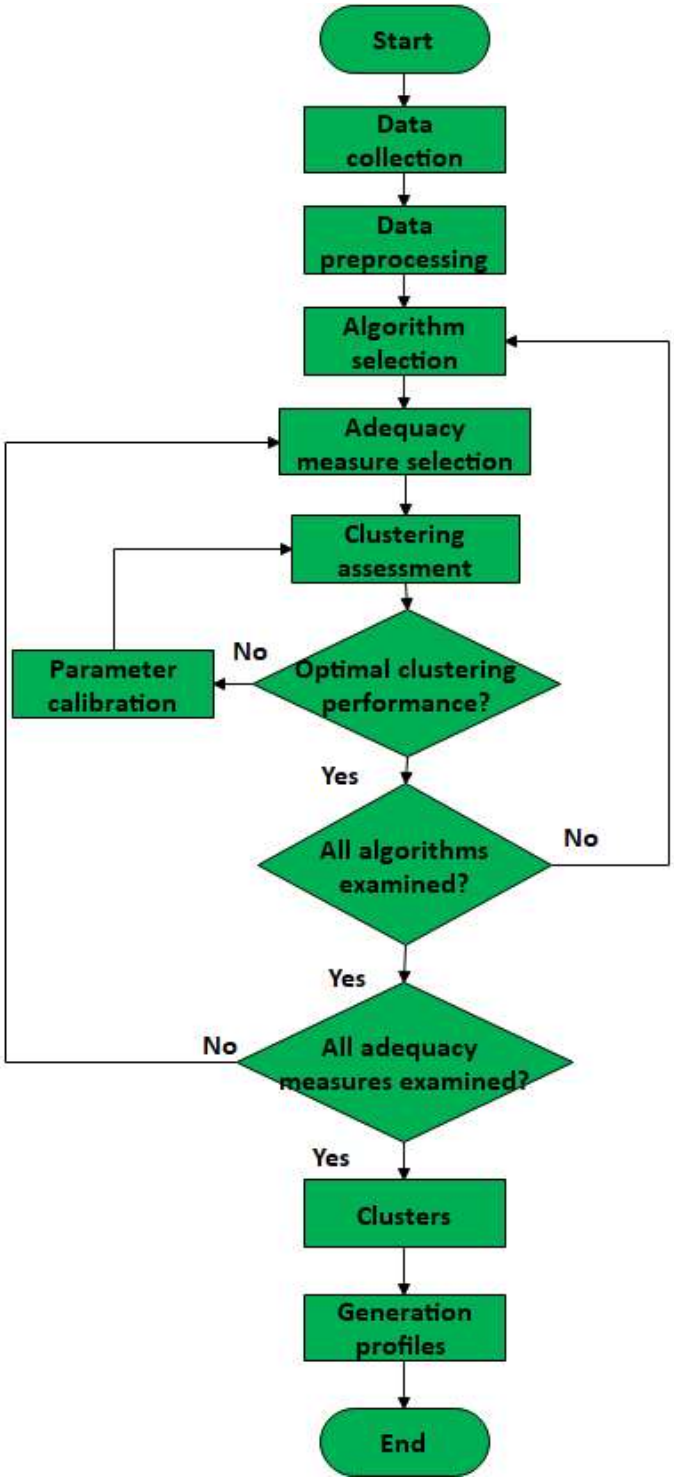


Figure 2. Flow-chart of the PV generation profiling methodology.

$$C_k \neq \emptyset, \quad k = 1,2, \dots K \tag{4}$$

$$C_i \cap C_j = \emptyset, \quad i = 1,2,\dots,K,\ j = 1,2,\dots,K,\ i \neq j \tag{5}$$

$$\bigcup_1^K C_k = X \tag{6}$$

Clustering assigns each pattern to one cluster such that:

$$u_{km} = \begin{cases} 1, & x_m \in C_k \\ 0, & x_m \notin C_k \end{cases} \tag{7}$$

Equation (3) determines the centroids of the clusters. These are expressed in per unit values, i.e., values within the [0,1] range. By multiplying with p_{max} the centroids are expressed in physical units. Let s_k be an indicator denoting the static profile of k-th cluster. Static profiles are derived utilizing historical PV generation data. They can be used to describe a data set P in cases where no new data records are available. Static profiles are obtained through the clustering of the historical data. It is:

$$s_k = c_k\ p_{max} \tag{8}$$

The set that contains the static profiles is denoted as $S_k = \{s_k, k = 1,\dots,K, 1 \leq K \leq M\}$. Dynamic profiles are derived using weather-related information. A linear regression model is applied per cluster to derive the relationship between generation and weather variables.

2.2 STATIC PROFILES FORMULATION

2.2.1 Outlier detection

The term “outlier” refers to atypical patterns that are present in the original data set. These patterns can be real extreme values of the target variable or metering failures and erroneous operation of the metering equipment. The outlier detection is held via the DBSCAN algorithm. Let n be a point for clustering. The DBSCAN algorithm considers three types of points, namely, central points, density-reachable points or outliers based on the following rules:

1. A point n is a central point if at least *MinPts* points are at a distance ε from this and these points are directly accessible from n. No point is accessible from a non-central point.
2. A point o is densely accessible from a point n, if there is a path $n_1, \dots, n_n$ with $n_1 = n$ and $n_n = o$ where every n_{i+1} is directly accessible from n_i. This means that all points of the paths must be central, except for o.
3. Points that are not accessible from anywhere else are considered outliers.

If the point n is a central point, then it forms a cluster together with all the points (central or not), which are accessible from it. Each cluster contains at least one central point. Parameter *MinPts* the minimum number of points required to create an area, i.e., a cluster. The algorithm starts from a random point. Next all the points of the neighborhood are obtained, i.e., points that correspond to greater distances than ε. If the number of points is greater than *MinPts*, a cluster is created. If the opposite is the case, the point is considered as an outlier. If a point is a dense part of a cluster, then its neighborhood is a subcluster of it. Thus, all points within the neighborhood are added to the cluster, as well as points in the ε-neighborhood of each of these points. This process continues until

finding the densely connected cluster. Then a new point is selected, and the above procedure is followed, to detect additional cluster or noise.

2.2.2 Clustering Algorithms

The selection of clustering algorithm is critical to the methodology`s robustness. The methodology involves the family of hierarchical agglomerative clustering algorithms. Hierarchical clustering groups data not in one step as in partitioning clustering, but through a number of steps. It refers to the creation of sub-clusters within clusters organized in the form of a tree, i.e., dendrogram. The tree structure consists of a) nodes in its trunk that correspond to clusters, b) leaves corresponding to sub-clusters and c) root that corresponds to the initial number of clusters. The initial number of clusters equals the number of patterns. This means that initially every pattern in the data set consists of singleton clusters, i.e., clusters that contain one pattern. The root node of the tree represents the entire set of patterns and each leaf node is considered a pattern. Intermediate nodes describe the degree to which patterns are close together, whereas the height of the tree usually expresses the distance between two patterns or clusters or between a pattern and a cluster. The results of the clustering can be obtained by "cutting" the tree at different levels. Through a series of agglomerations, one cluster is produced that includes all patterns. The user provides the termination condition of the agglomerations and the clusters are obtained. A characteristic of hierarchical clustering is the fact that it assigns patterns to the various clusters permanently. Figure 3 shows an example of a dendrogram. The horizontal axis refers to the observations, i.e., the patterns. The vertical axis refers to the distortion, which is a clustering adequacy measure. Hierarchical agglomerative clustering algorithms start with M singleton clusters. Next, the calculation of the MxM takes place. The minimum distance between the centroids is calculated and they are merged to form a new cluster. The matrix is updated, and the distances are calculated again. The process is stopped when one cluster is derived. The merge of a pair of centroids depends on the distance function between the current centroid and the new one.

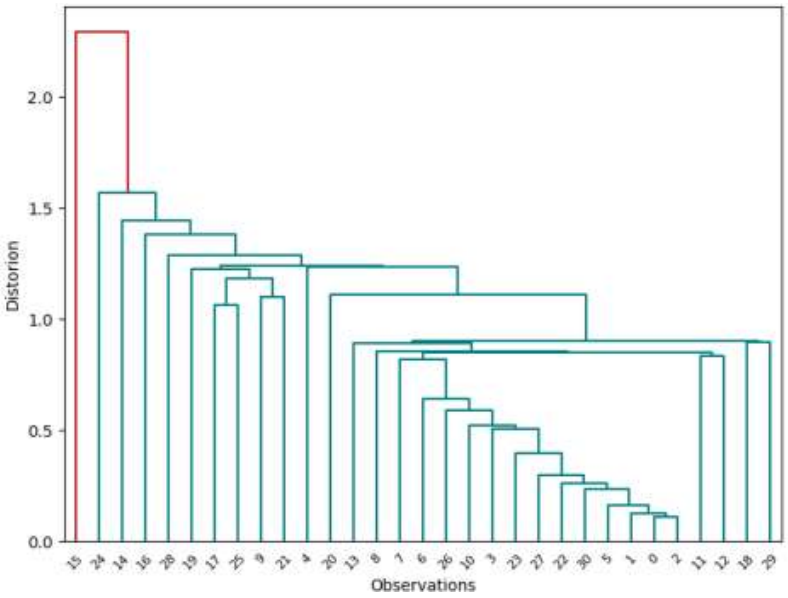


Figure 3. Example of a dendrogram.

Let $d(\cdot,\cdot)$ be the operator that denotes distance. Let C_m, C_l be two different random singleton clusters of the data set X. Also, let C_i, C_j be the clusters that are merged to form a new cluster C_{ij}. The minimum distance $d(C_i, C_j)$ is calculated as:

$$d(C_i, C_j) = \min_{1 \le m,l \le M} d(C_m, C_l) \quad (9)$$

Equation (9) provides the minimum distance between all the random pairs C_m, C_l.
After the calculation of the distance the proximity matrix is updated by calculating the distances between the newly formed cluster C_{ij} and the rest. The general form of the distance function is given by:

$$d(C_l, (C_i, C_j)) = a_i d(C_l, C_i) + a_j d(C_l, C_j) + \beta d(C_i, C_j) + \gamma \left| d(C_l, C_i) - d(C_l, C_j) \right| \quad (10)$$

where a_i, a_j, β and γ are coefficients that define the agglomerative algorithm. Let M_i, M_j and M_l be the number of patterns that belong to clusters C_i, C_j and C_l, respectively. Table 1 present the values of the coefficient per algorithm. There are seven agglomerative algorithms that differ in the definition of the distance function. The following simplified terms can be used to name the algorithms: Single (SL), Complete (CL), Average (UPGMA), Weighted (WPGMA), Ward (MVM), Median (WPGMC) and Centroid (UPGMC) [2].

Table 1. Coefficients of the hierarchical agglomerative algorithms.

Algorithm	a_i	a_j	β	γ
SL	0.50	0.50	0	0.50
CL	0.50	0.50	0	0.50
UPGMA	0.50	$\frac{M_j}{M_i + M_j}$	0	0
WPGMA	0.50	0.50	0	0
WPGMC	0.50	0.50	-0.25	0
UPGMC	$\frac{M_j}{M_i + M_j}$	$\frac{M_j}{M_i + M_j}$	$\frac{M_i M_j}{(M_i + M_j)^2}$	0
MVM	$\frac{M_i + M_l}{M_i + M_j + M_l}$	$\frac{M_j + M_l}{M_i + M_j + M_l}$	$\frac{-M_l}{M_i + M_j + M_l}$	0

2.2.3 Clustering Evaluation

Clustering evaluation is held with a set of clustering validity indices or adequacy measures. Each measure examines specific characteristics of the generated partitions, like compactness, separation, and others. The measures are built on the Euclidean distance. Let x_s and x_t be two patterns that $(x_s, x_t) \in X$. The Euclidean distance $d_{Eucl}(\cdot,\cdot)$ is calculated as:

$$d_{Eucl}(x_s, x_t) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_s^t - x_t^t)^2} \quad (11)$$

Let S_k be the subset of X that belong to cluster C_k. The DI is the geometric mean of the inner-distances between the members of the subset S_k. It is:

$$\mathrm{DI} = \sqrt{\frac{1}{2M_k}\sum x_k \in S_k d_{Eucl}^2(x_k, x_m)} \quad (12)$$

The CI refers to the ratio of the dispersion among the different clusters and the dispersion within the same cluster:

$$\mathrm{CI} = \frac{M-K}{K-1}\cdot\frac{\sum_{k=1}^{K} M_k \cdot (c_k - p)\cdot(c_k - p)^t}{\sum_{m.k=1}^{K}\sum_{x_m \in X}(x_k - c_k)\cdot(x_k - c_k)^t} \quad (13)$$

where p is the arithmetic mean of X. The DBI is expressed as the ratio of the sum of the most similar clusters to the distance of their centroids:

$$\mathrm{DBI} = \frac{1}{K}\sum_{s,t=1}^{K}\max_{s\neq t}\left\{\frac{d(S_s)+d(S_t)}{d(c_s, c_t)}\right\} \quad (14)$$

where $c_s, c_t \in C_K$. The adequacy measures operation is two-fold: a) To evaluate an algorithm`s performance and b) to denote the optimal number of clusters [3].

2.3 DYNAMIC PROFILES FORMULATION

The static profiles are the product of the clustering process. Through the exploitation of the static profiles, conclusions about the PV generation patterns can be obtained. Also, no continuous and intense collection of further PV generation are needed. The static profiles result in lesser requirements of data collection since the generation patterns can be described and represented with the profiles. However, if it is required, additional generation data can be gathered and a new clustering analysis will be executed and new static profiles will be formed. In cases where limited data are available generation data are available, the dynamic profiles concept can be adapted. The dynamic profiles correspond to the updated static profiles. This update refers to the adjustment of the static profiles considering recent recordings or predictions of temperature and irradiation. This

means that the shapes of the static profiles are adjusted to fit to the recent meters of meteorological data or predictions of them. With this approach, the need of continuous gathering of generation data is limited. The profiles are updated only with weather related data. Therefore, the dynamic profiles are weather adjusted profiles of the daily PV generation curve clusters. After the application of clustering, the patterns are grouped into clusters. Apart from generation data, solar irradiation and temperature curves can be grouped together to the clusters that the corresponding days grouped to. A piece-wise linear regression model is applied to each cluster whose regression parameters are estimated using a search algorithm. The number of regression models equals the number of clusters. Hence, the number of dynamic profiles equals the number of static profiles.

The scope is to derive the function that describes the relationship between generation and one hand, irradiation and temperature on the other. Through the Gradient Descent search algorithm, the best possible statistical fit to the historical generation data is extracted. The linear regression approach considers that the dependent variable, i.e., generation is a linear function of the explanatory or independent variables. Let d_k be the dynamic profile of the k-th cluster. Also, let T_k and I_k be the temperature and solar irradiation of the k-th cluster, respectively. The model to derive the dynamic profile is:

$$d_k = \beta + x_1 T_k + x_2 I_k + e \tag{15}$$

where β is the intercept term, x_1 and x_2 are random variables and e is the error term. The random variables are estimated via the Ordinary Least Squares Algorithm (OLSA). Equation (15) denotes that the dynamic generation profile is a linear combination of the weather variables that affect the generation capacity of the PV system. In order to increase the forecasting accuracy, the OLSA minimizes error e:

$$e = \sum_{i=1}^{K} d_k - \hat{d}_k \tag{16}$$

where $\hat{d}_k$ is the predicted dynamic profile if the external variables T_k and I_k where forecasts. Equation (15) can be expressed in matrix form as:

$$d_k = \begin{vmatrix} d_k^1 \\ \vdots \\ d_k^{\mathrm{m}} \end{vmatrix}, X = \begin{bmatrix} 1 & x_{1,1} & \cdots & m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m,1} & \cdots & x_{m,m} \end{bmatrix}, \beta = \begin{vmatrix} \beta_0 \\ \vdots \\ \beta_m \end{vmatrix}, e = \begin{vmatrix} e_1 \\ \vdots \\ e_m \end{vmatrix}$$

Equation (16) can be rewritten as:

$$e = Y^T Y - Y^T X\beta - \beta X^T Y + \beta^T X^T X\beta \tag{17}$$

The error minimization is accomplished via the following condition:

$$\frac{\partial e}{\partial \beta} = 0 = \beta^T X^T X - Y^T X \tag{18}$$

The solution of the above equation is:

$$\beta = (X^T X)^{-1} X^T Y \tag{19}$$

The evaluation of the performance of the regression model, the following indices are utilized:

- Coefficient of correlation or R^2:

$$R^2 = 1 - \frac{RSS}{SYY} \tag{20}$$

where RSS is the residual sum of squares, which is expressed as:

$$RSS = e^2 \tag{21}$$

and SYY is the sum of squares, which is expressed as:

$$SYY = \sum (d_k - \bar{d}_k)^2 \tag{22}$$

where $\bar{d}_k$ is the average of d_k.

- Adjusted coefficient of correlation, R^2_{adj}:

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(k-1)}{p-k-1} \right] \tag{23}$$

where $p$is the number of variables excluding the constant term β.

- F-statistic:

$$F = \left[\frac{(\tilde{x}_1^2) \sum (x_k - \bar{x})^2}{\tilde{\sigma}^2} \right] \tag{24}$$

where $x_1, \ldots, x_2$ are the independent variables of the regression model and σ is the variance [4]. According to Equation (15) there are two independent variables.

3 RESULTS

Indicatively, this section presents some results on clustering of a PV data set. The PV system is 10 kWp and located in Northern Greece. The period of the data is 01/09/2018-30/09/2019 and the metering interval is 15 min. The MVM or Ward`s algorithm is selected. The algorithm is executed for 2 to 34 clusters. For each number, the values of the measures are checked. Figure 4 presents the shapes of the metrics per number of clusters. The vertical axis in number 8 denotes that this number is optimal number. The CI and DI are decreasing as the number of cluster increases. For 8 clusters, the following values are obtained: CI=170.69, DBI=1.46 and DI=246.64. Table 2 presents the day type distribution over the 8 clusters. Figure 5 shows the centroids of the clusters. Figure 6 and Figure 7 display the patterns that belong to clusters 2 and 7, respectively.

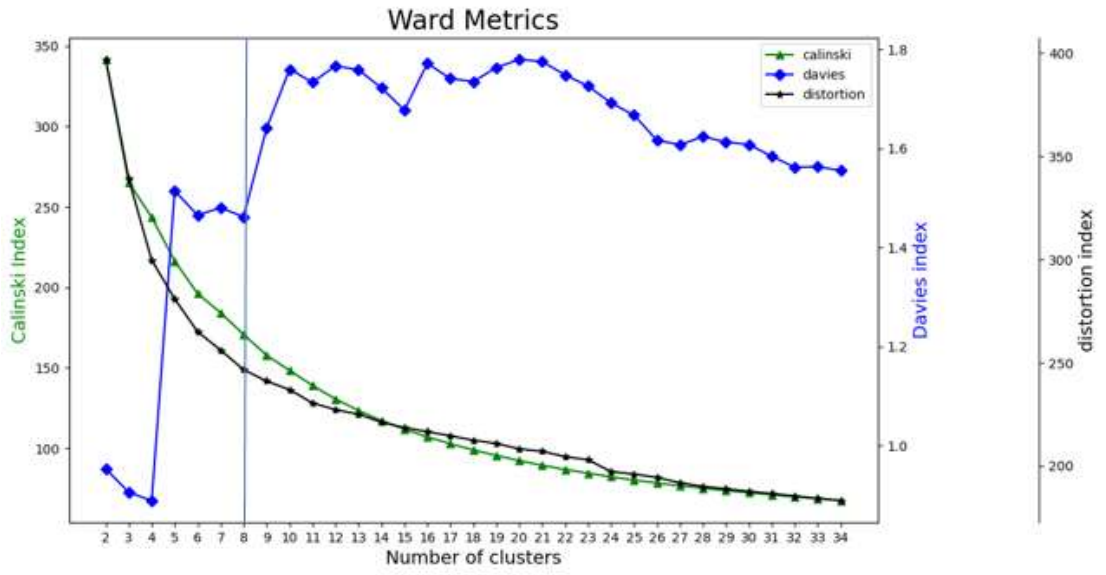


Figure 4. Adequacy measures per number of clusters.

Table 2. Day type distribution.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
January	23	2	5	1				
February		2	14	12				
March	1	1	6	21	1			1
April			7		7	3	5	8
May			1		11	2	5	12
June		1			14	1		14
July			1		3	15		12
August						25	1	5
September						21	5	4
October	2		11	18				
November	6	11	13					
December	6	10	15					

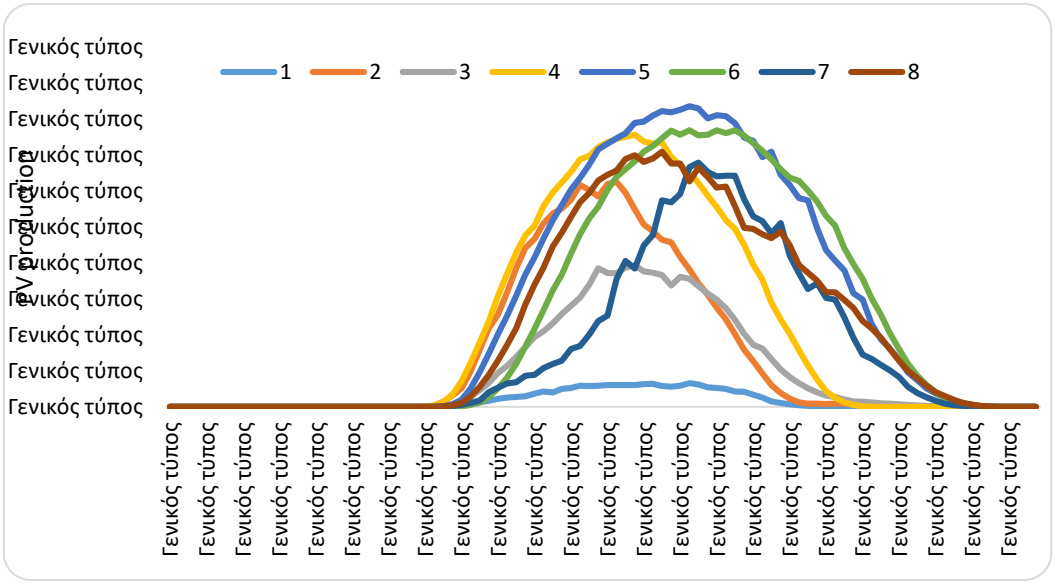


Figure 5. Centroids of the clusters.

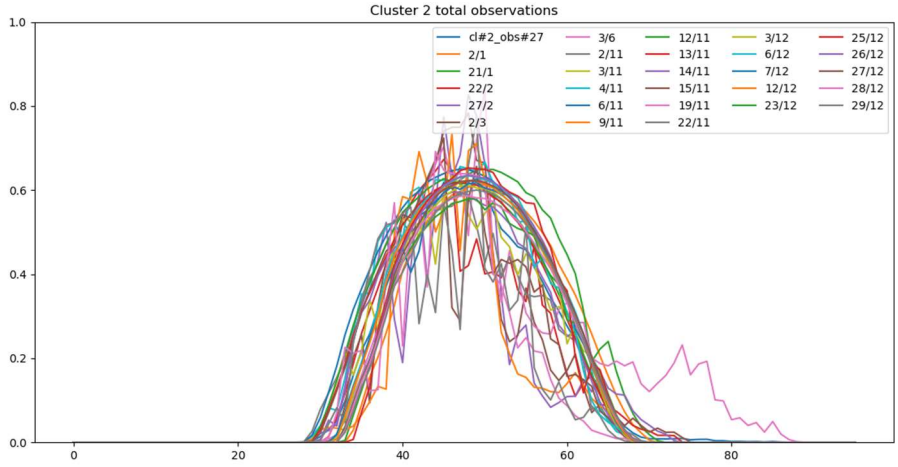


Figure 6. The patterns of Cluster 2.

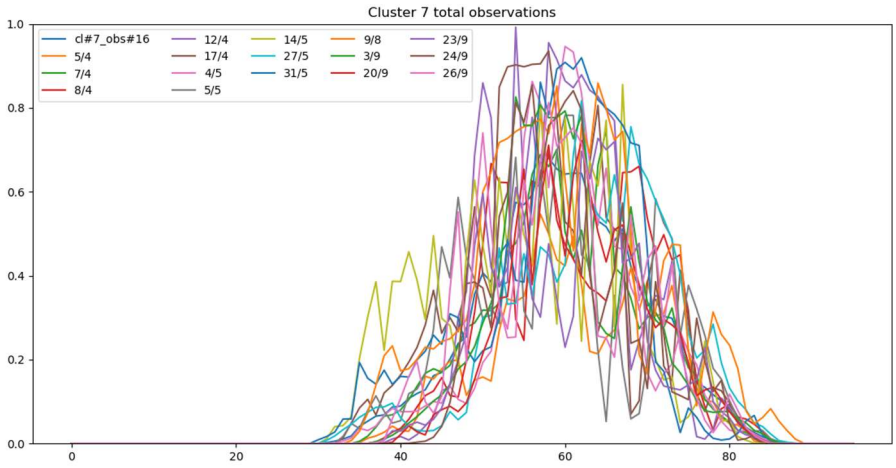


Figure 7. The patterns of Cluster 7.

4 GUI DESCRIPTION

The Graphical User Interface (GUI) has been developed in Python™ programming language [5]. Specifically, we have utilized the Python™ 3.7 version and the "tkinter" package, which is built into the Python™ standard library. Figure 8 shows the schematic representation of the GUI`s operation. The GUI can be distributed as an independent .exe file. The user can install the GUI package without installing other Python™ related software. Also, no prior Python™ programming expertise is needed to implement and work with the GUI.

In the first stage, the user is asked to enter the PV generation data. The data can be entered in MS Excel™ format. Actually, ".xlsx" or ".csv" types can be used. The data can be entered in row or column vectors. Also, the data should be expressed in physical units, i.e., kW, MW and others. The GUI supports different time resolution of the recorded data, namely 1 min, 5 min, 15 min and 60 min. The data are normalized in the [0,1] range by dividing each data entry with the maximum value of the specific data set. The data should contain only the PV generation. The user should manually enter the start date that the recordings refer to. This is needed as to connect dates to the data recordings. The clustering outcomes includes the distribution of the dates into the clusters. The next stage includes the selection of the clustering algorithm. There are seven available algorithms. Afterwards, the user is asked to select the adequacy measure that will validate the clustering results. There are three available measures. Also, the user can select all measures and display them in a single figure.

The following stage deals with the number of clusters. The desired number of clusters should be between two and the number of days that the data cover. For instance, if one year of data is provided, the maximum permitted number of clusters is 365. The DBSCAN clustering algorithm is used to track outliers. The DBSCAN threshold parameter is supplied by the user. When clustering is executed, the values of the measures are displayed starting from two and up to the number of clusters provided by the user in the previous stage. Also, the clusters are displayed.

The user can select a cluster and examine the daily PV generation curves that are grouped together in the specific cluster. Also, through the DBSCAN algorithm the outliers are displayed and marked with red color. The user can decide whether they will be removed from the data or otherwise. In the case, the user wishes to remove them, a second clustering takes place that will cluster the new data set with one or more outliers removed. Again, all algorithms and adequacy measures are available for this new clustering execution. The clusters centroids refer to the static PV generation profiles.

A further analysis of the data set involves the static profiles update using external variables. Specifically, the user is asked to select the profiles that wishes to update. Next, the user provides solar irradiation and temperature data. A linear regression is applied to the selected cluster in order to extract the relationship between the generation and external variables. The dynamic profile of the cluster is the output of the regression model. Though this approach, the profiles can be updated using newly information of the external variables, i.e., newly recorded data or forecasts.

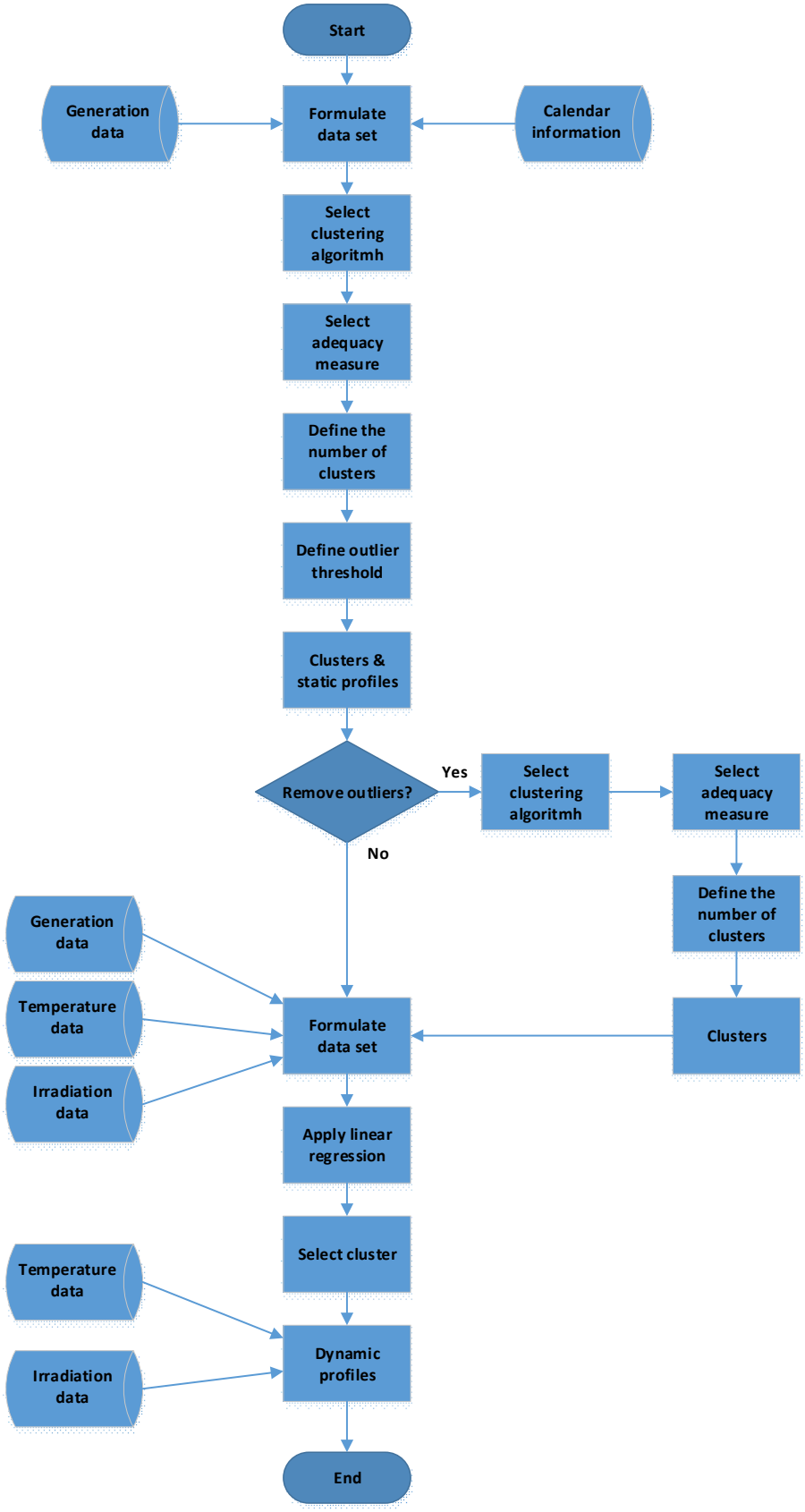
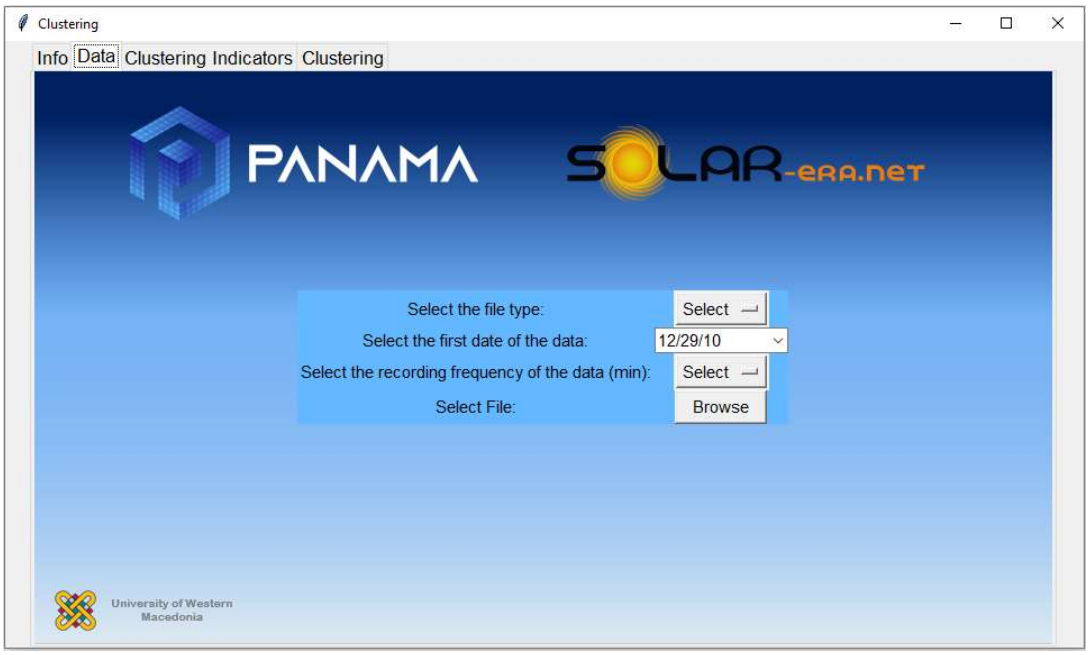


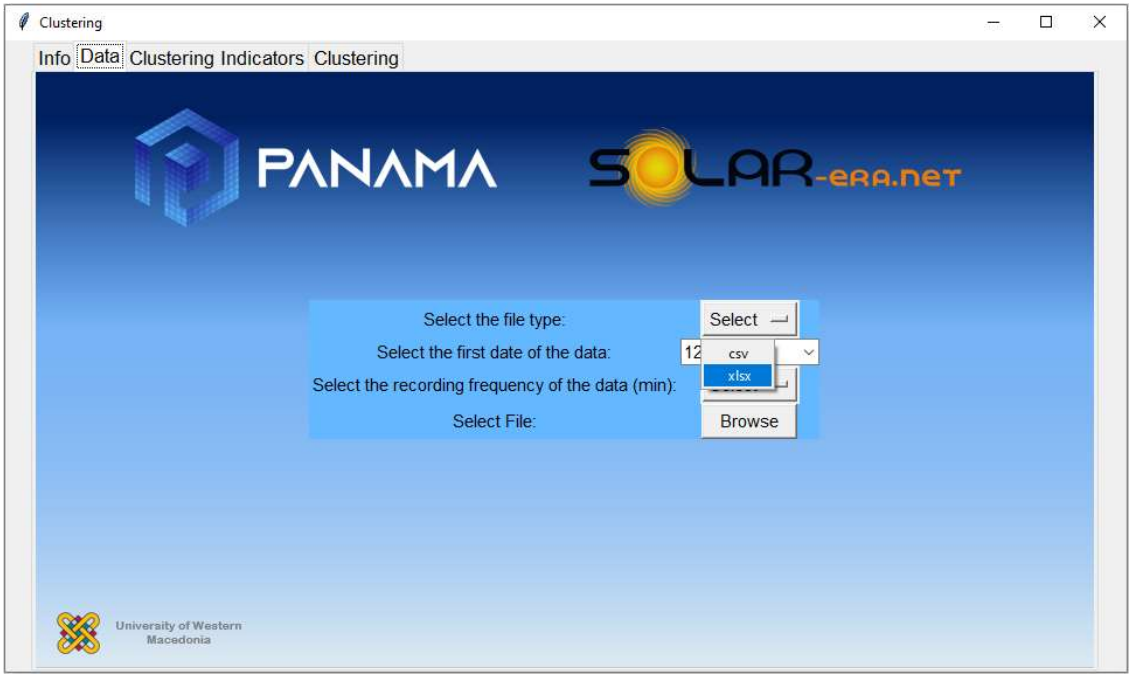
Figure 8. Flow-chart of the GUI`s operation.

The GUI consists of four tabs (Info, Data, Clustering Indicators, Clustering). The first tab ***"Info"*** contains the logos of the project.

As we move on to the **"*Data*"** tab the user is called to import the file with the data and give some additional information.

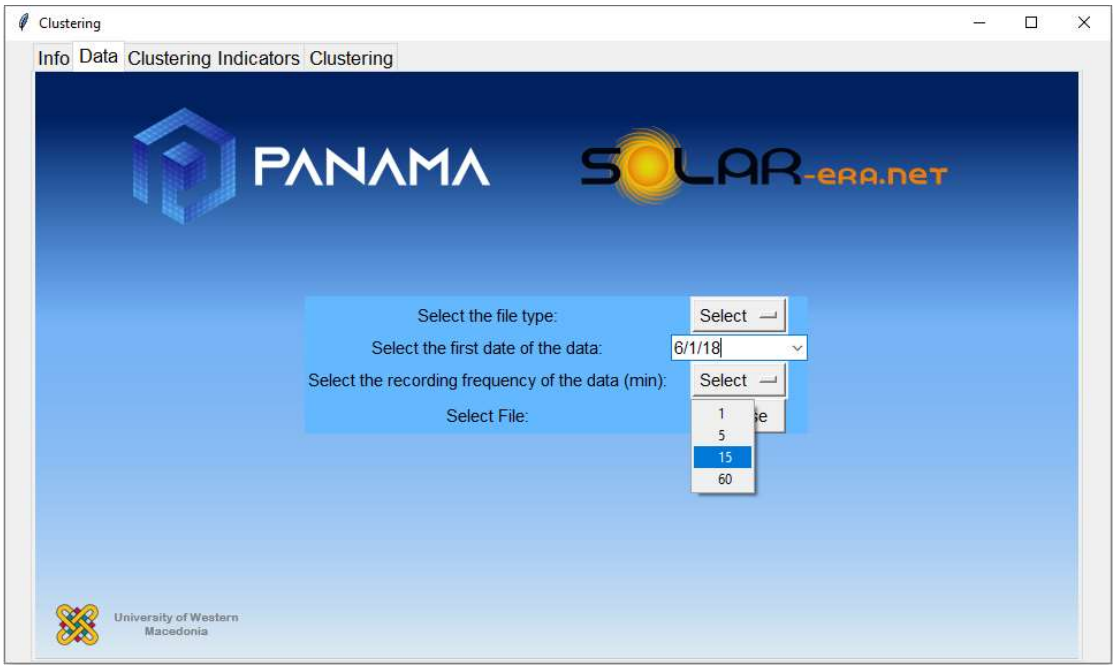


At the first field **"*Select file type:"*** the user should select the type of the file from the option menu. Specifically, the GUI supports two different types: a) CSV files and b) XLSX files. The imported data should be given as row or column data.



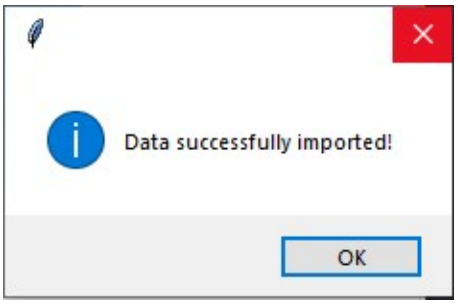
Accordingly, the user should insert the first date of the data. For instance, if the 1st date of the data is the 1st of June of 2018 the user has to select it from the calendar.

Afterwards, the user has to select the recording frequency of the data. This is important in order to separate the data into days, since the observations of each day varies according to the recording frequency. The GUI supports four different recording periods: a) 1 min, b) 5 min, c) 15 min and d) 60 min.



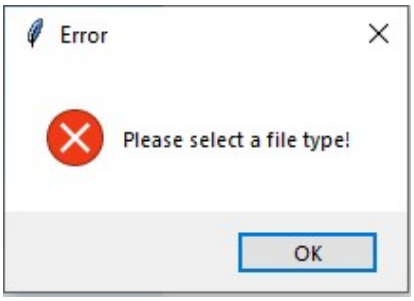
Finally, the user, by pressing **"*Browse*"**, can browse, select the data file from the computer and open it.

If the data are imported successfully in the GUI the following message will appear.

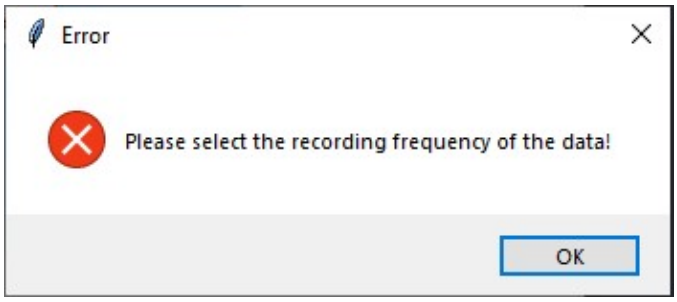


The GUI also provides error messages:

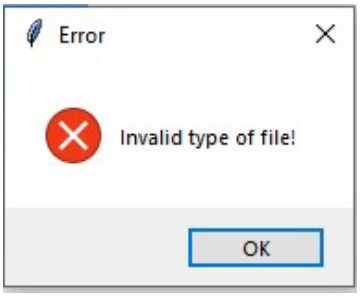
1. If the user does not select a file type.



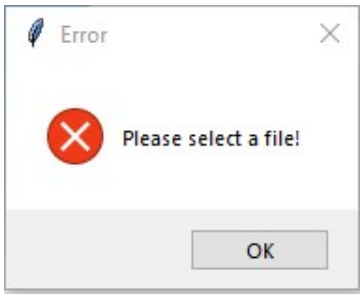
2. If the user does not select a recording frequency.



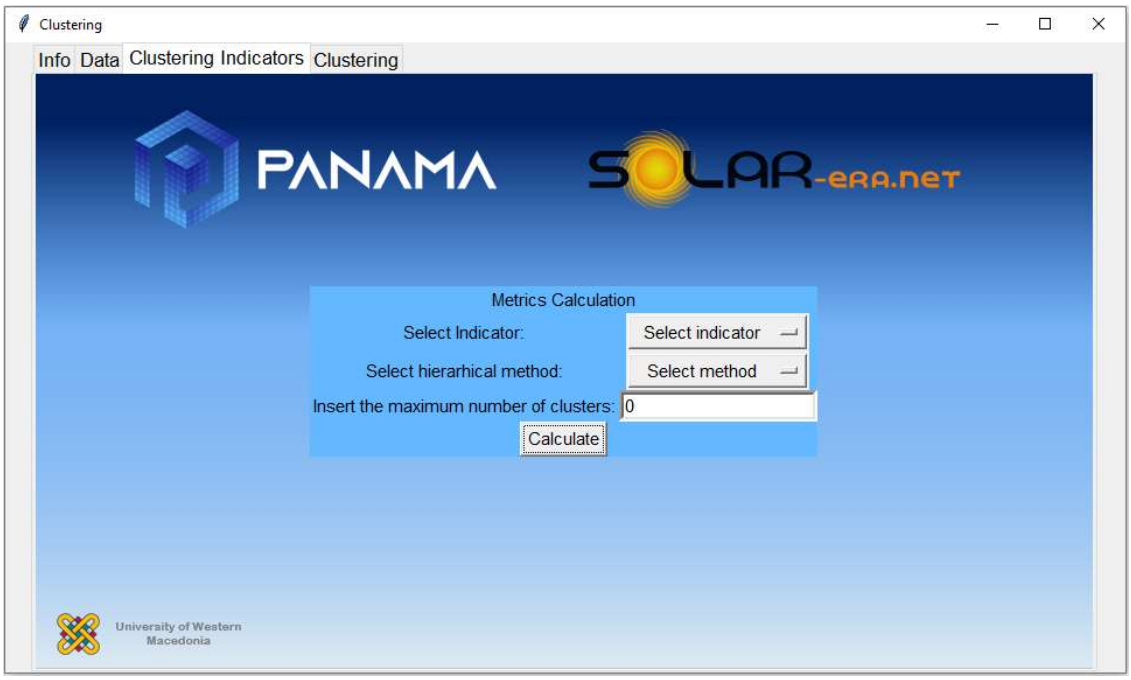
3. If the user does not select a correct type of file from the browse window.



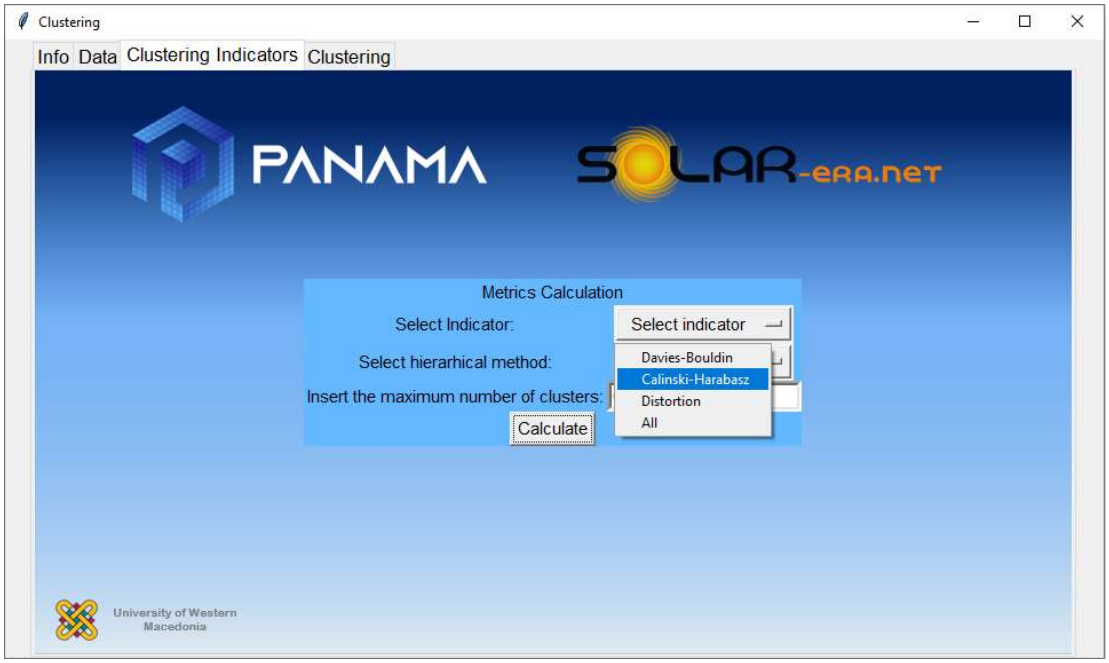
4. If the user closes the browse window without selecting a file.



After the user imports the data he can move on to the tab **"*Clustering indicators*"**.



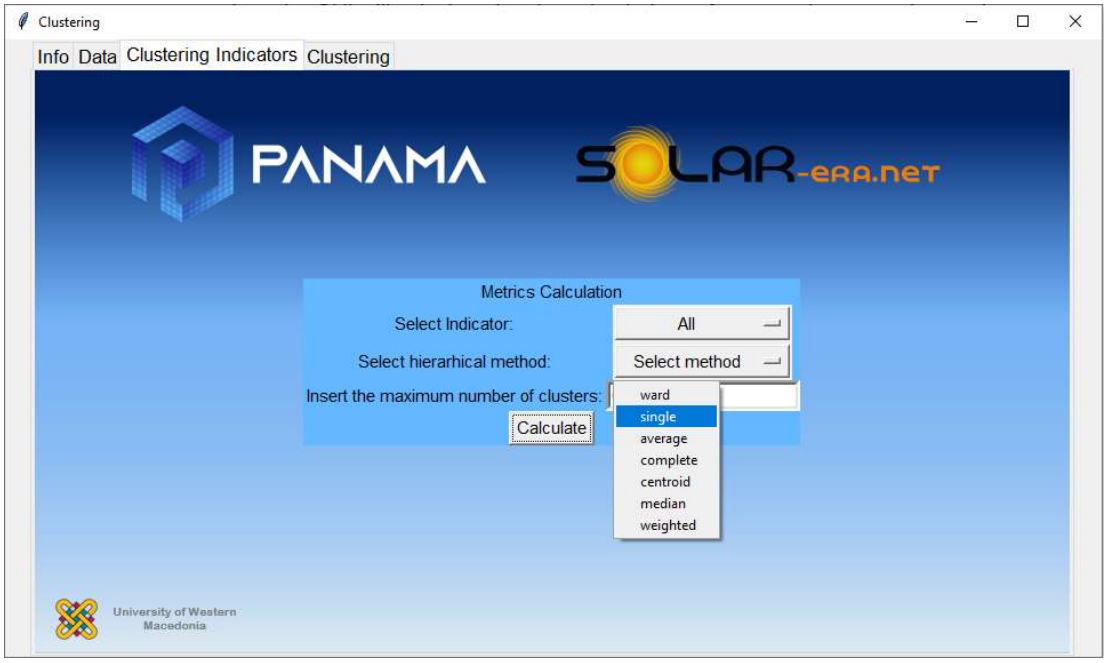
This tab can help the user to plot the three indicators in respect to the number of clusters. Specifically, the user should select from the option menu which indicator he wants to plot. He has four different options: a) ***"Davies-Bouldin"***, b) ***"Calinski-Harabaz"***, c) ***"Distortion"*** and d) ***"All of them"***.



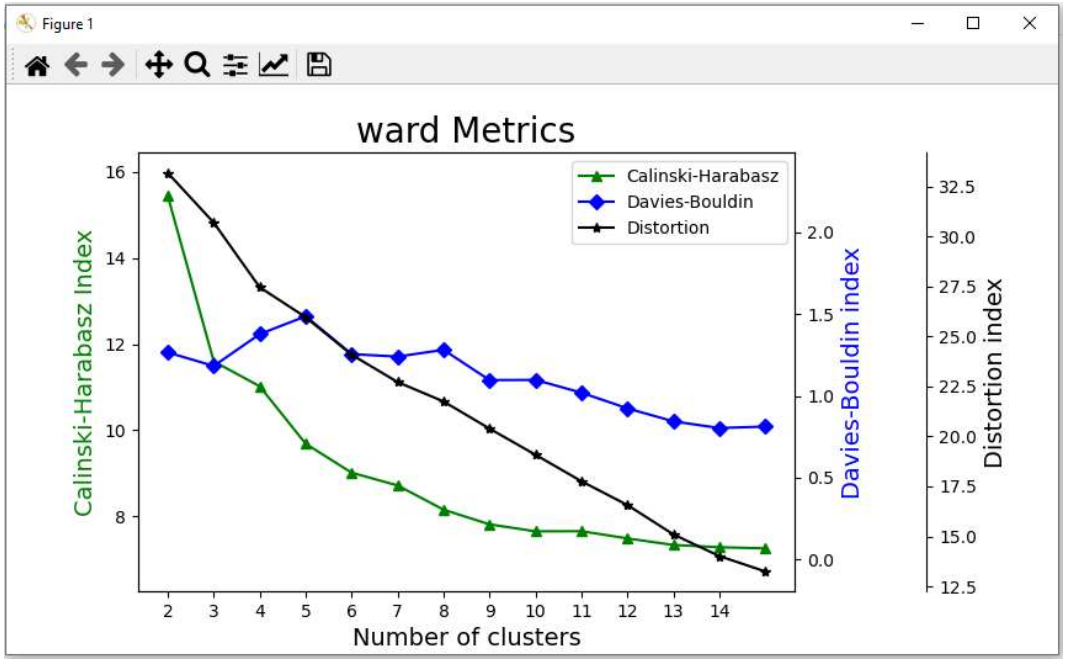
Accordingly, the user should select a hierarchical clustering method. The GUI give as options all the seven hierarchical clustering algorithms:

1. ***"ward"***
2. ***"single"***

3. ***"average"***
4. ***"complete"***
5. ***"centroid"***
6. ***"median"***
7. ***"weighted"***

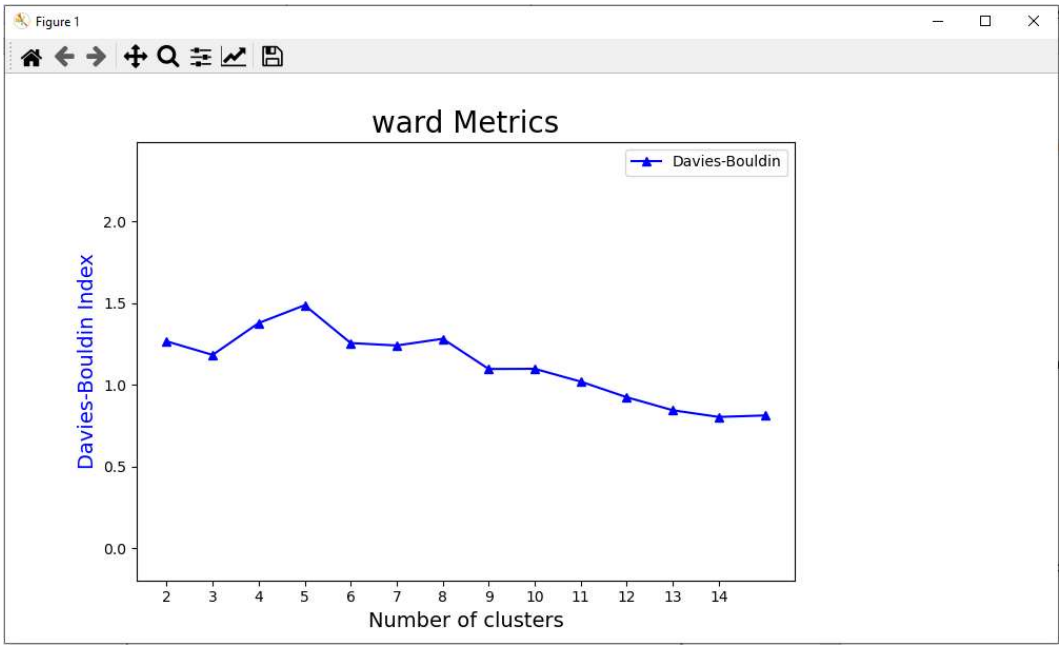


After the method's selection the user should give the maximum number of clusters. This is necessary since the GUI will calculate the clustering indexes from two clusters to the maximum number of clusters. For example, the user selects ***"All"*** for the indicators plot and the ***"ward"*** hierarchical method. Also, he sets the maximum number of clusters equal to 15. When he presses the ***"Calculate"*** button the following window will appear. The name of the selected method is included at the top of the plot.

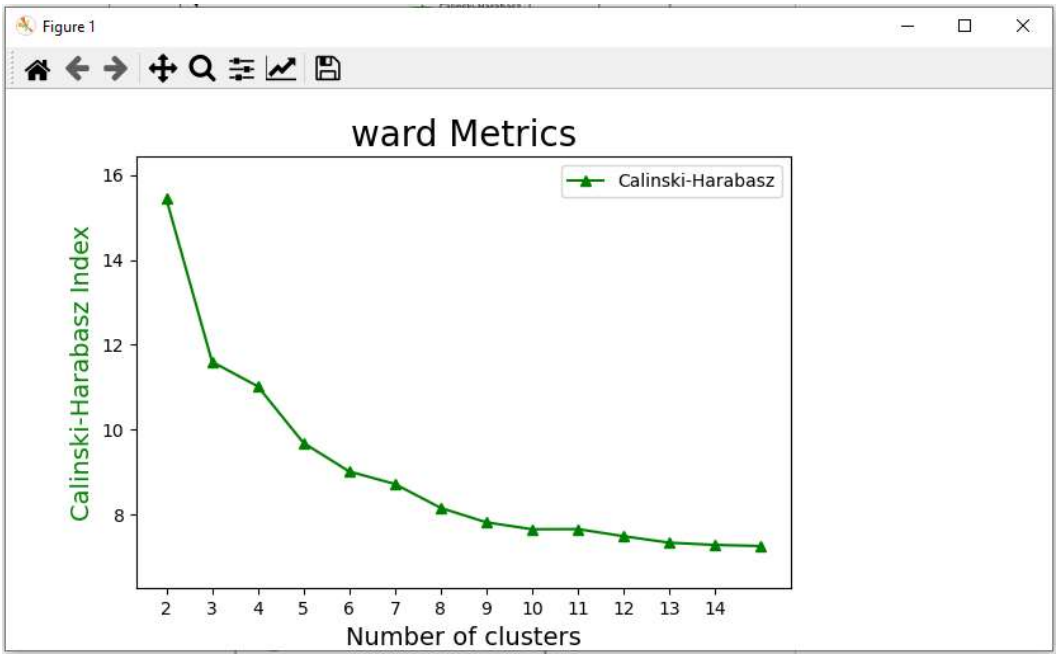


The following figures represents the widow with the plot for the rest of the indicators.

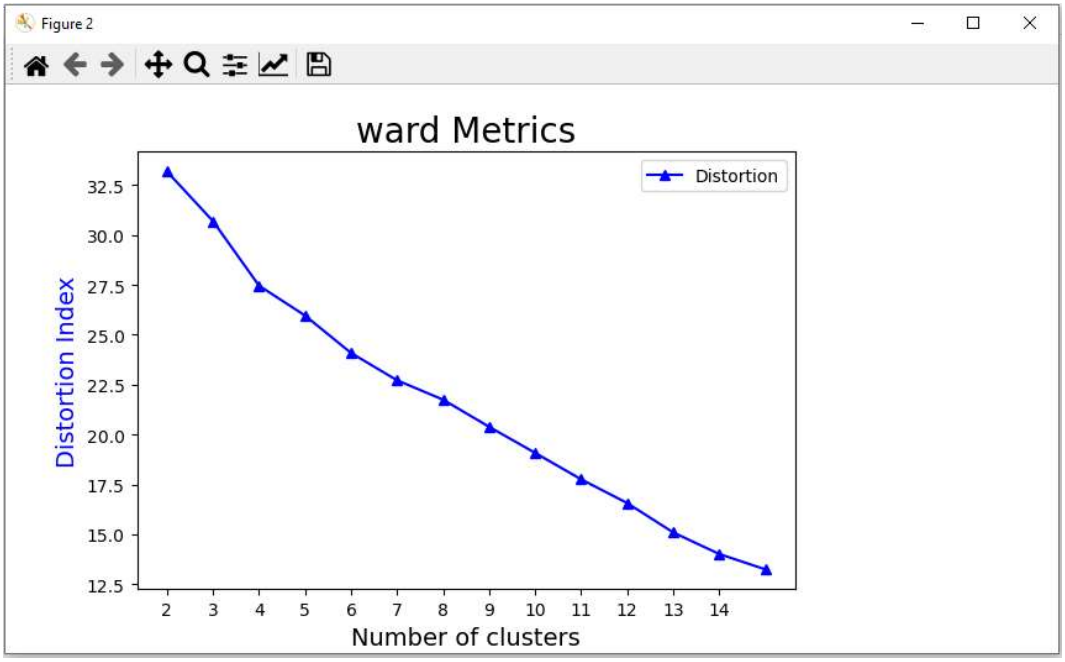
1. Davies -Bouldin



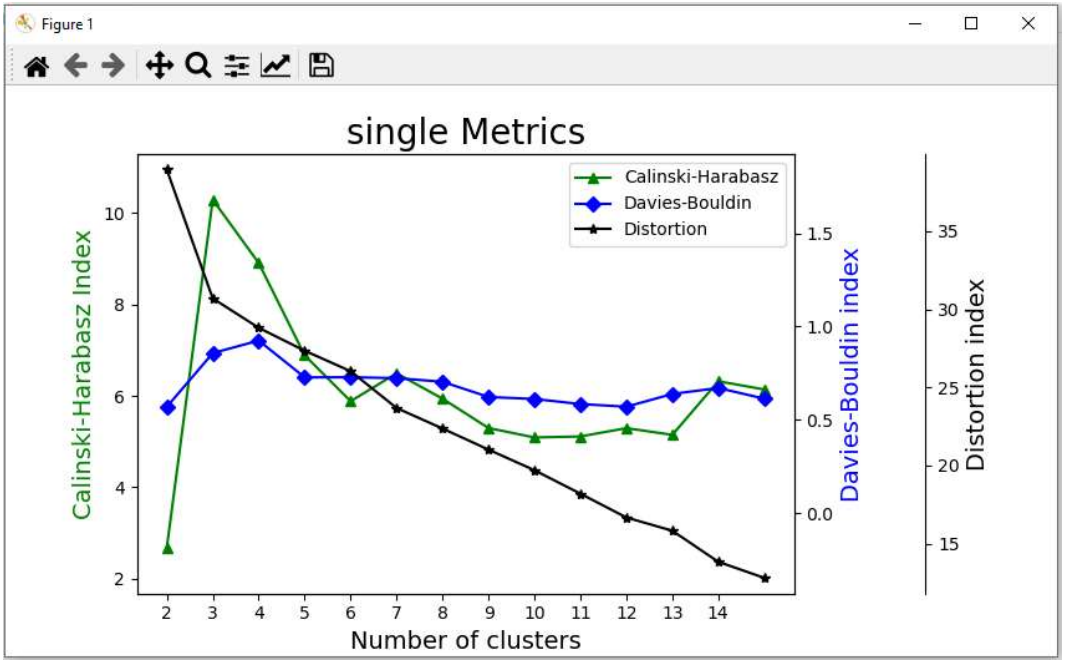
2. Calinski-Harabasz



3. Distortion

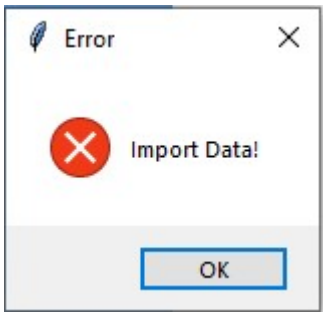


The figure below illustrates the plot for all clustering indicators when the single method is selected.

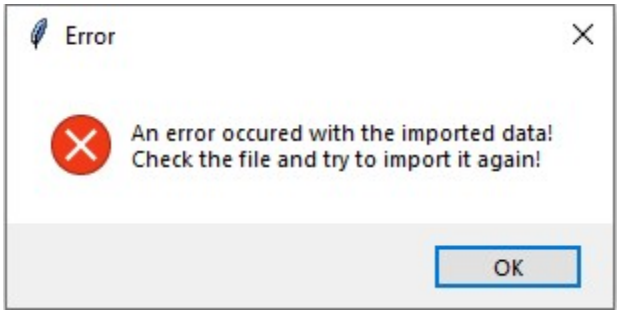


At the second tab five different errors can appear.

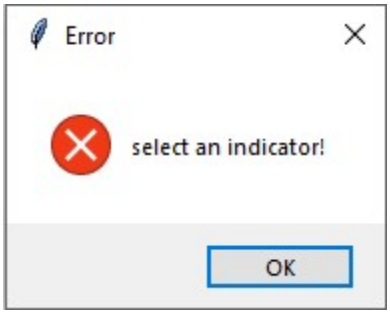
1. If the user moves to the tab ***"Clustering Indicators"*** without import file.



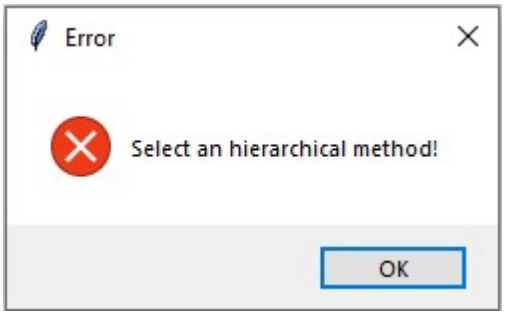
2. If the user imports invalid data



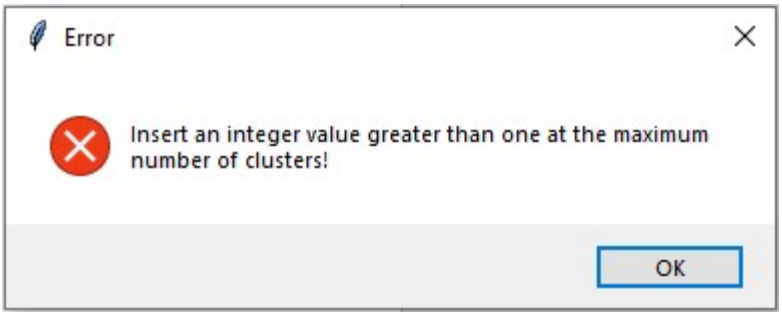
3. If the user does not select indicator.



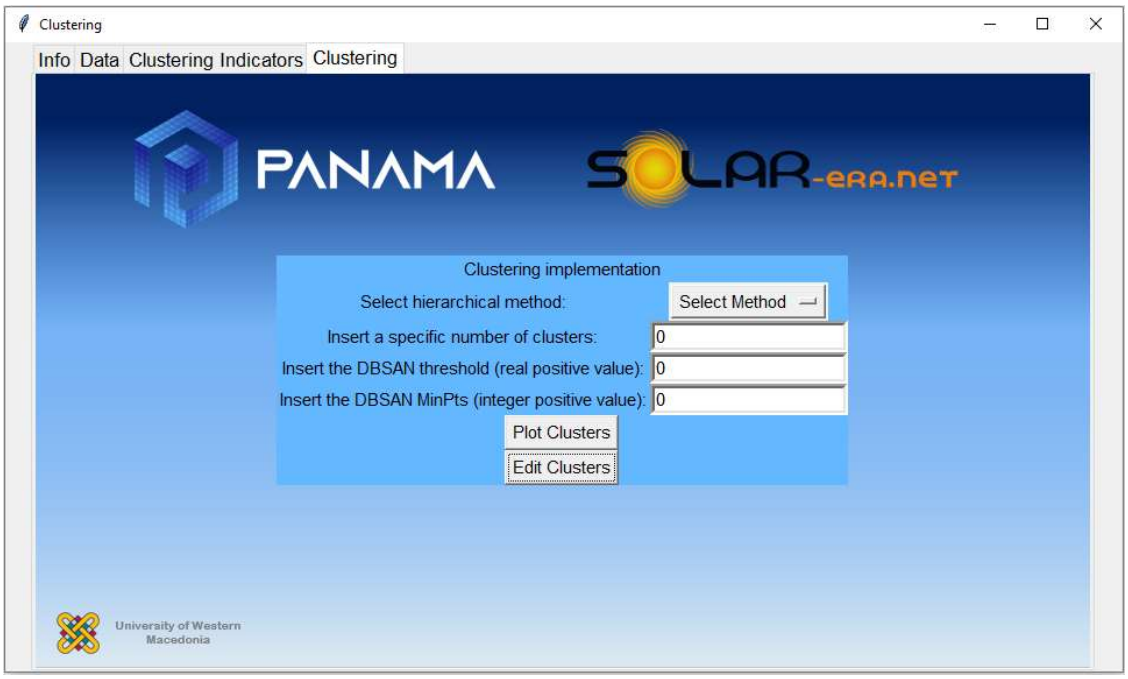
4. If the user does not select clustering method



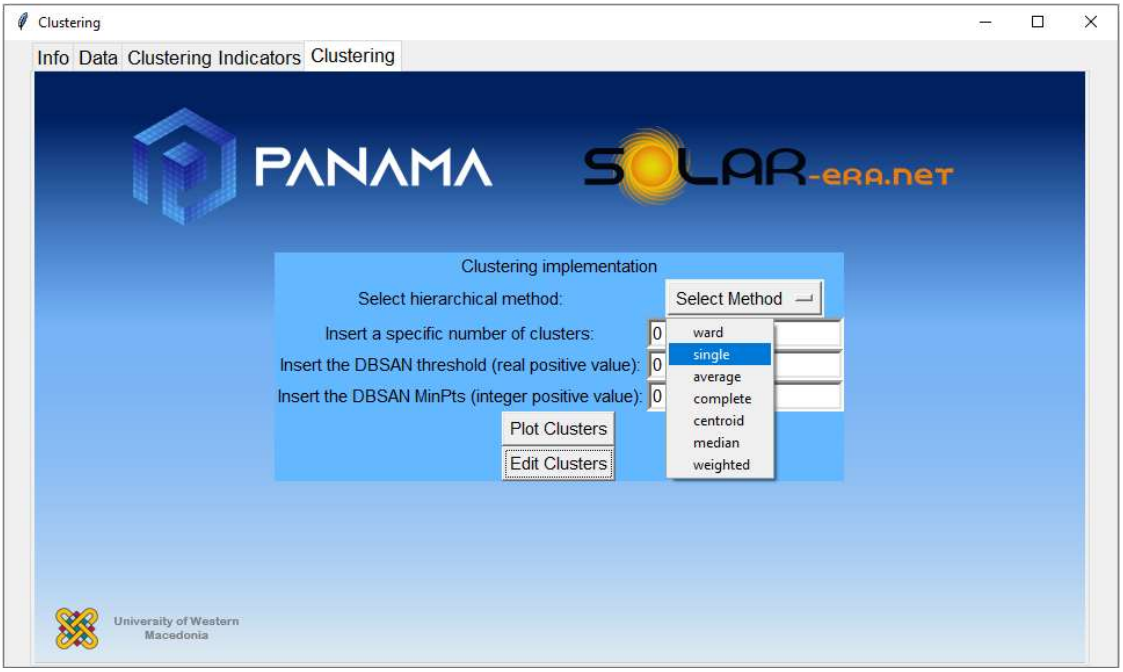
5. If the user does not select an integer value greater than one at the field ***"Insert the maximum number of clusters".***



At the last tab the user is ready to implement the clustering method. The information obtained from the ***"Clustering Indicators"*** tab are valuable since the user determines the number of clusters according to them.

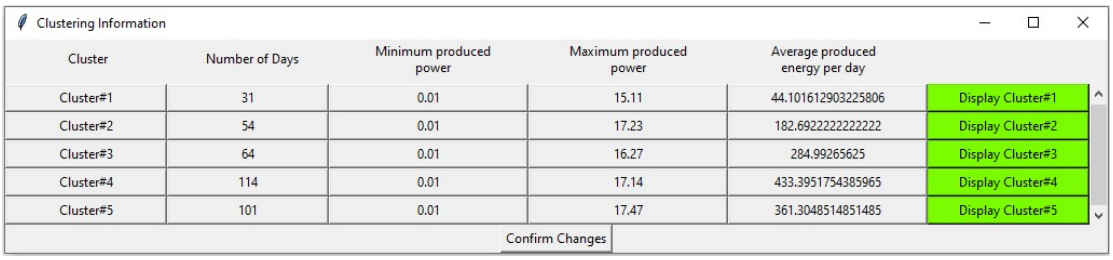


Before the user presses the ***"Edit Clusters"*** button he should define again a clustering method and select the specific number of clusters.



The user should also define a value for the *e* parameter and the MinPts parameter of the DBSCAN method. The DBSCAN clustering will be used later in the GUI to point out the invalid days within each cluster. The less the value of the DBSCAN threshold is set, the more invalid days the clustering method suggests.

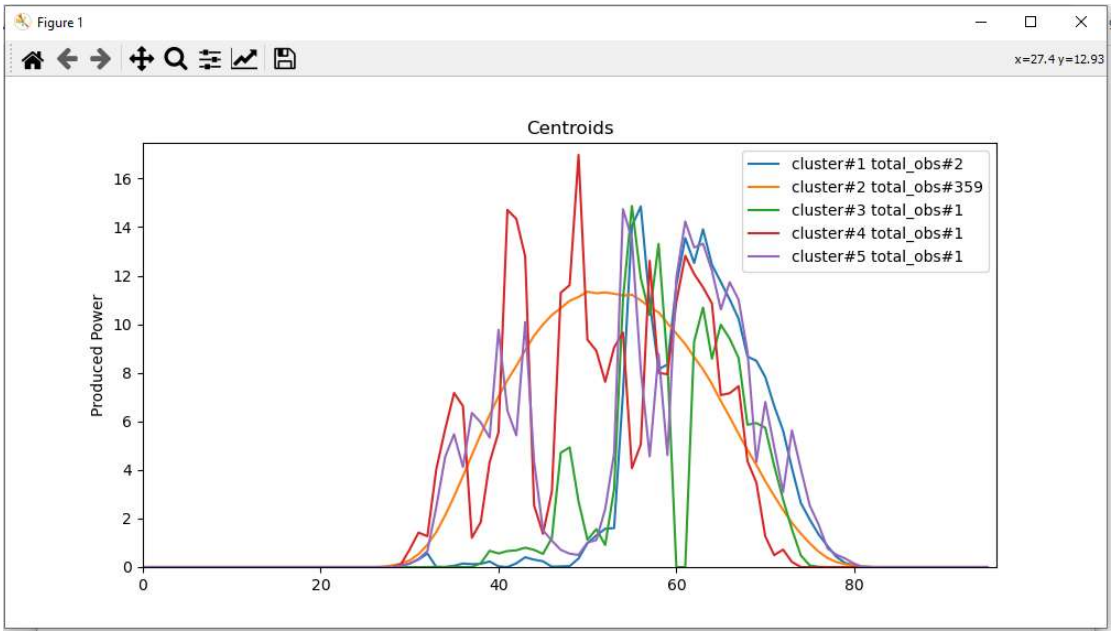
At first it is assumed that the user selects the ***"ward"*** method, defines the number of clusters equal to five, sets the DBSCAN threshold equal to 1.7 and defines the MinPts parameter equal to two. If the user presses the ***"Edit Clusters"*** button then an intra-cluster analysis is implemented with the DBSCAN clustering method, to detect the invalid days. The first column contains the name of the clusters and the second column contains the number of days each cluster has. Furthermore, the next three columns are used for a brief statistical analysis of the cluster and express the minimum and the maximum produced power, which is occurred in the cluster, as well as the average produced energy per day respectively. Finally, the last column contains colored buttons. The color of the buttons can be green, which indicates that all the days in the cluster are valid, red, which is used to warn the user that the cluster may contain invalid observations, and cyan, which is used for the singleton clusters. In the specific example all the buttons are green, meaning that no invalid day has been detected.



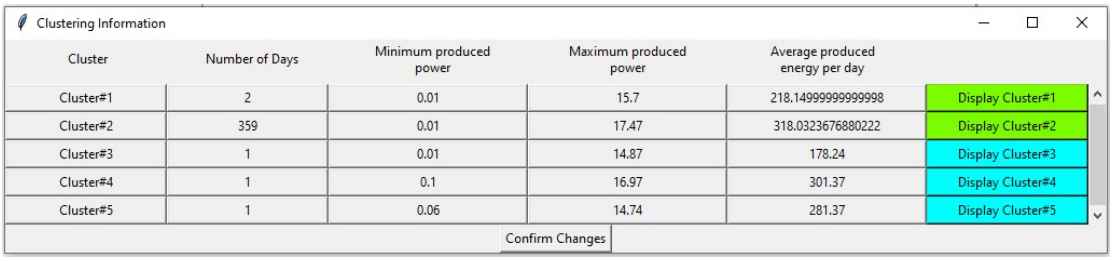
If the user wants to set another DBSCAN threshold he should close the window, without pressing the button ***"Confirm Changes"***, and insert another value.

In the following example, the user selects the ***"single"*** method, defines five number of clusters, sets the DBSCAN threshold equal to 1.4 and define the MinPts parameter equal to two. He can press the ***"Plot Clusters"*** button in order to plot the clusters. At figure with the cluster plot the

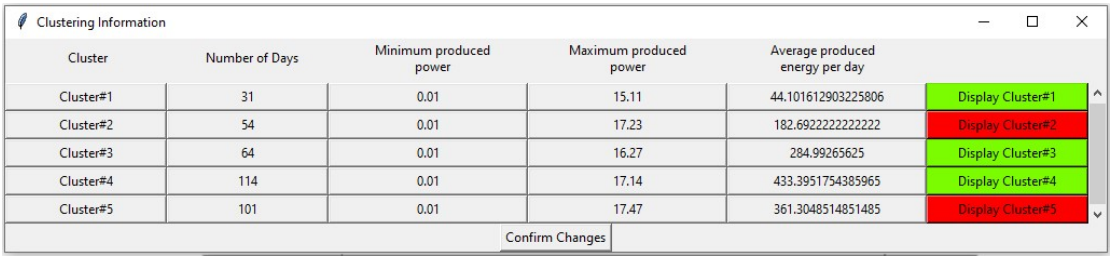
user can get general information about the centroids and the number of days, each cluster contains. Also, the y-axis represents the PV power and the unit of measurement depends on the user's data. At the specific example the data which are utilized are in values expressed in kW.



In order to get detailed information about the clusters the user can press the ***"Edit Clusters"*** button and a new window appears. In this case the DBSCAN defines three singleton clusters, with cyan color, and there are no invalid days within the two first clusters.



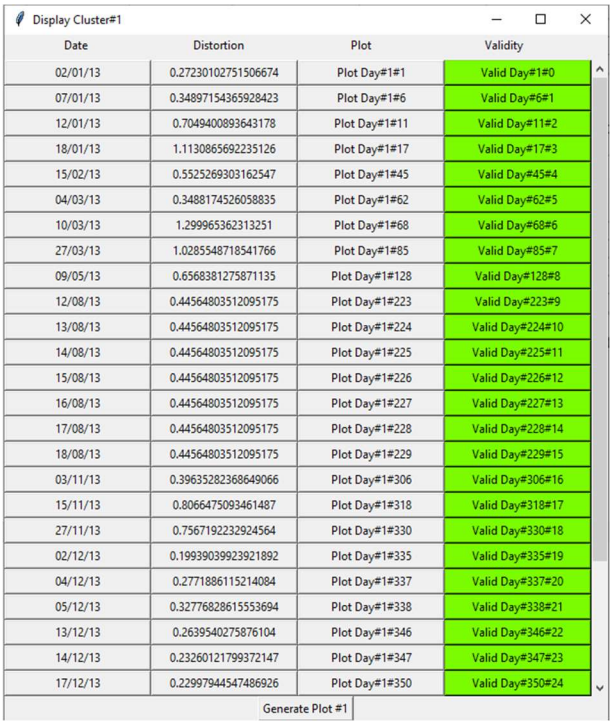
Since the red color is not included in the previous figure, we can change the clustering method to ward and keep the number of clusters and the DBSCAN threshold the same. In this case we can see that in the two out of the five clusters invalid days have been detected.



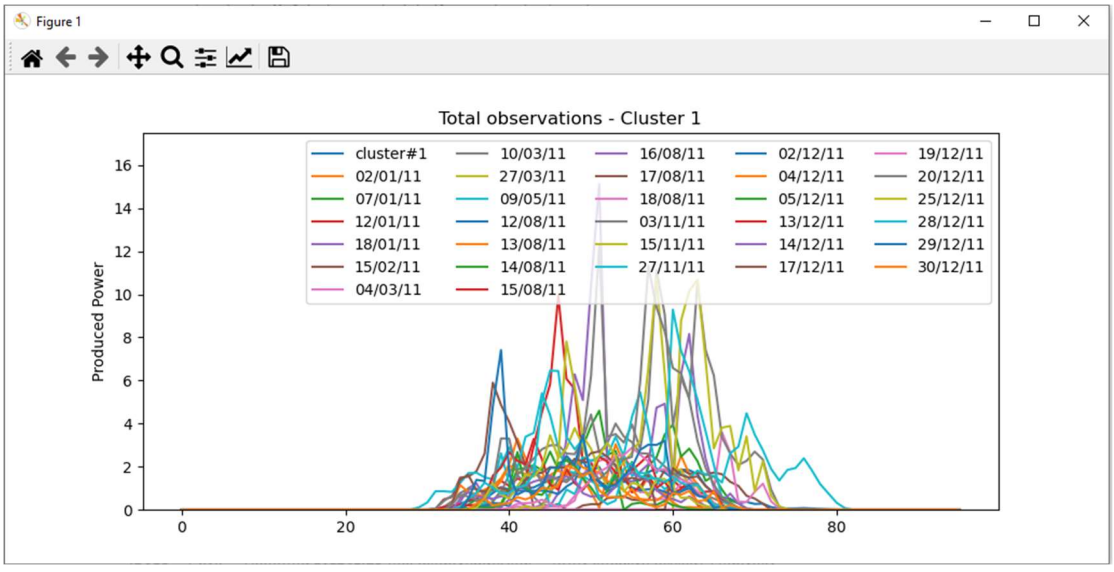
As it is obvious the selection of different clustering methods leads to different results and DBSCAN method detects different days as outliers of a cluster.

If the user presses the button ***"Display Cluster#1"*** a new window with an array will appear. As it is obvious the first column contains the dates of the days within the cluster. The second column

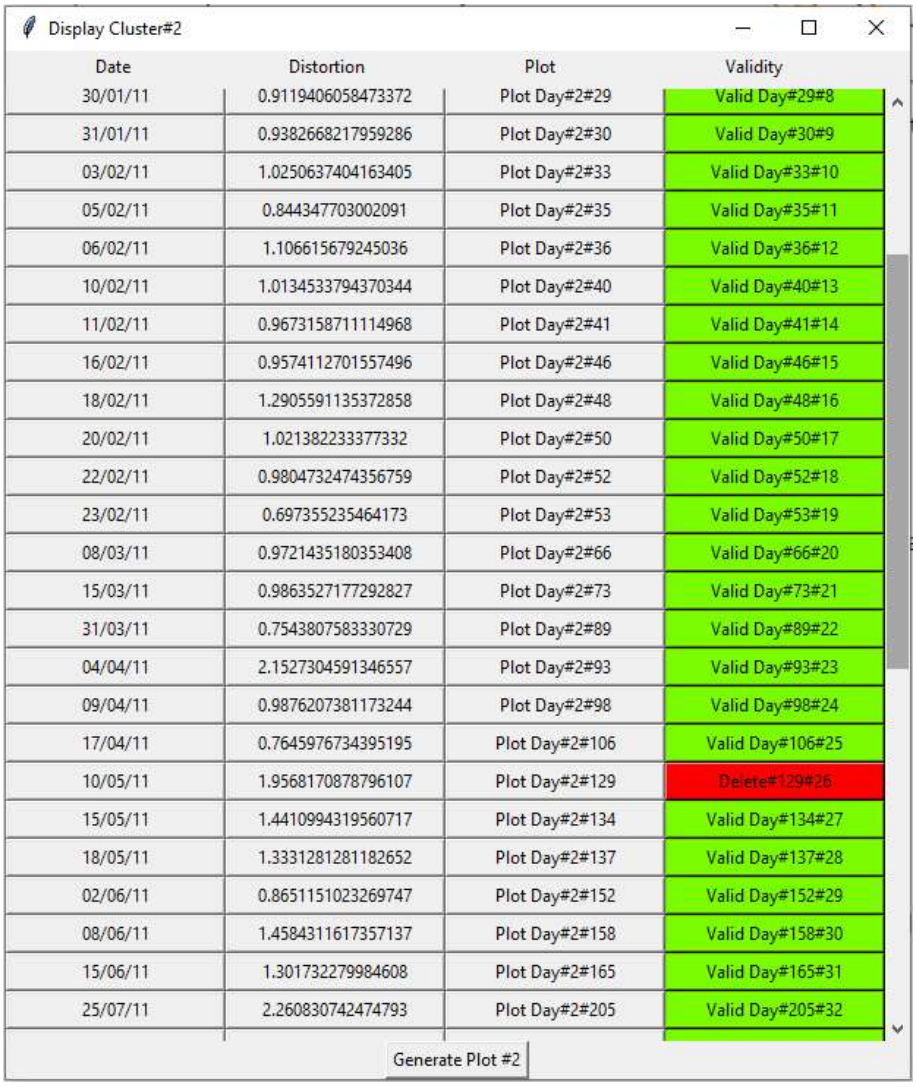
give information about the Euclidian distance between the PV production curve of the day and the centroid of the cluster. The buttons in the third column give the ability to the user to plot the specific day and the centroid of the cluster. In this way the user can decide whether the specific day is outlier or not. Finally, the buttons at the last column are used in order to delete a day from the cluster.



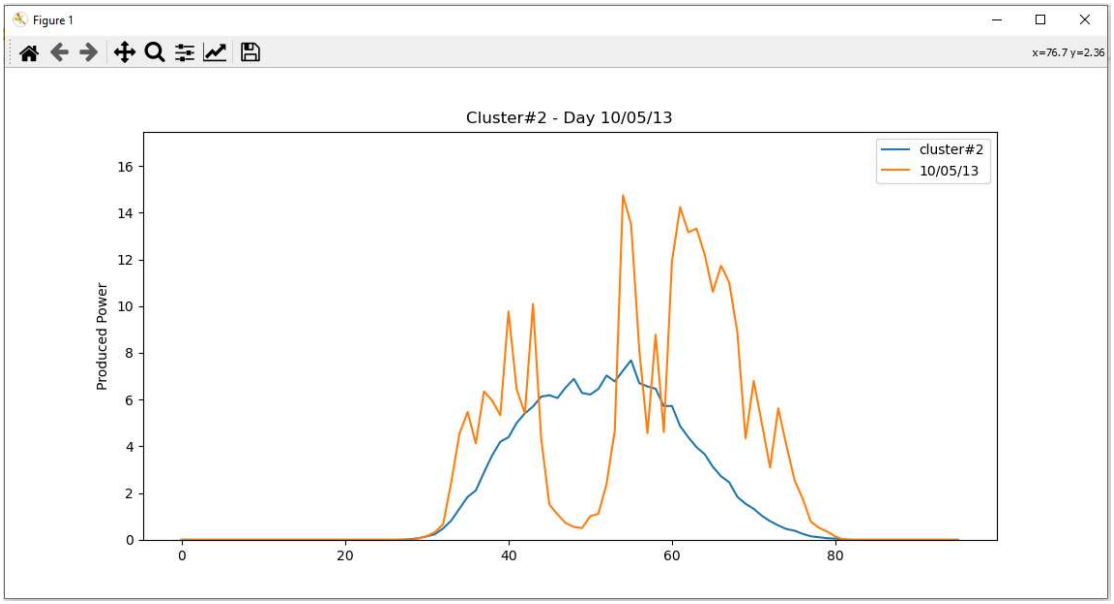
Moreover, the user can generate the plot of all the PV production curves within a cluster and the cluster's centroid respectively by pressing the ***"Generate Plot#1"***.



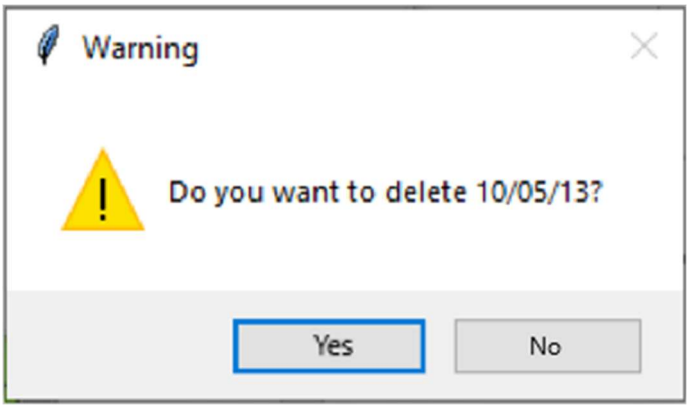
If the user selects to display all the information of the Cluster#2 the following window will appear.



The 10th of May is assumed to be an outlier for the specific cluster. If the user presses the ***"Plot Day#2#20"*** the plot will appear in a new window.



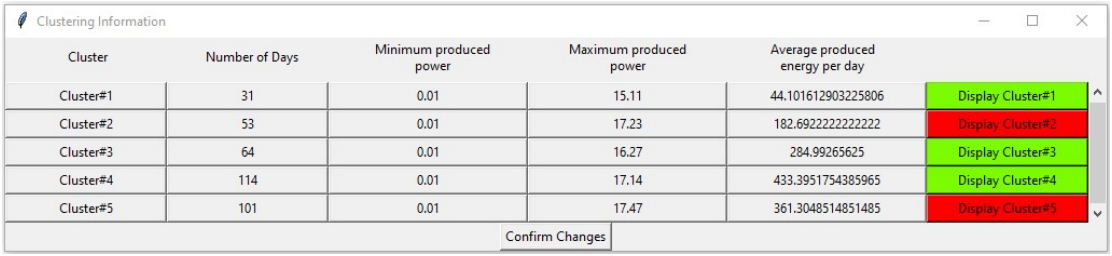
If the user agrees that the day is invalid then he should press the 'Delete#20#16'. The following window will appear in order to confirm or not the delete.



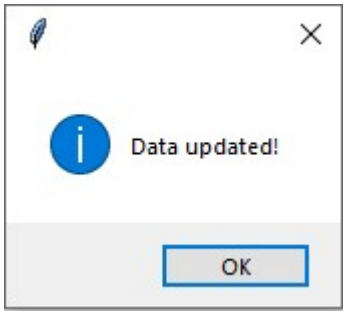
If the user presses ***"Yes"*** then the ***"Successfully Deleted"*** message will appear.

Also, at the ***"Display Cluster#2"*** window the specific day is assigned as ***"Deleted"***, as it is obvious from the figure bellow.

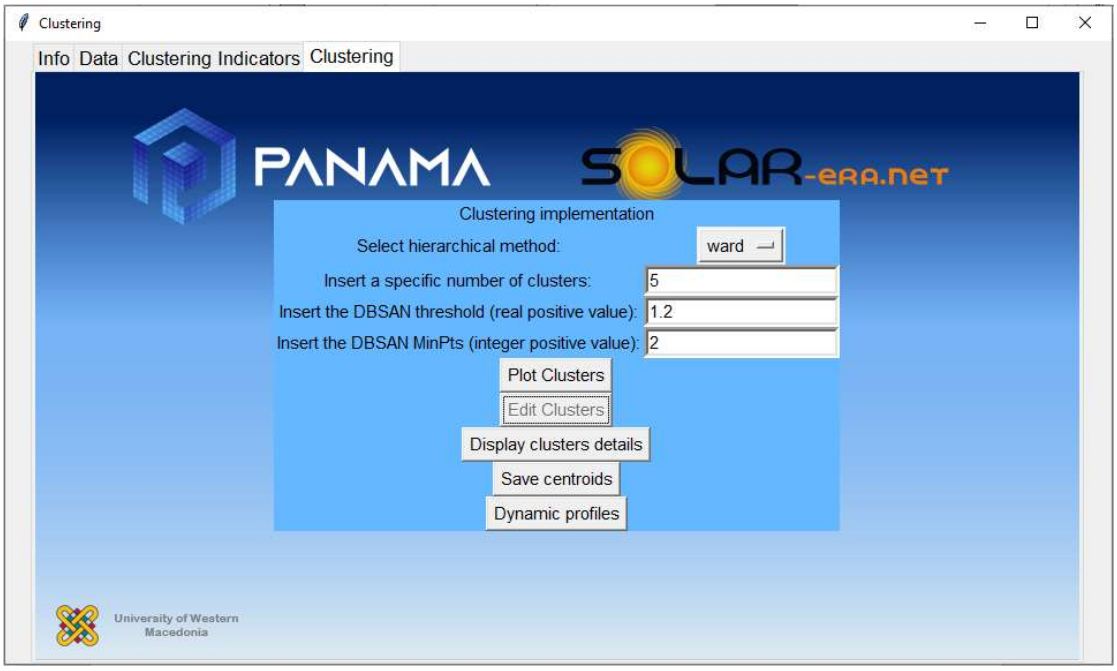
Moreover, if the user checks the window ***"Clustering Information"*** window the number of days in cluster 2 decreased by one.



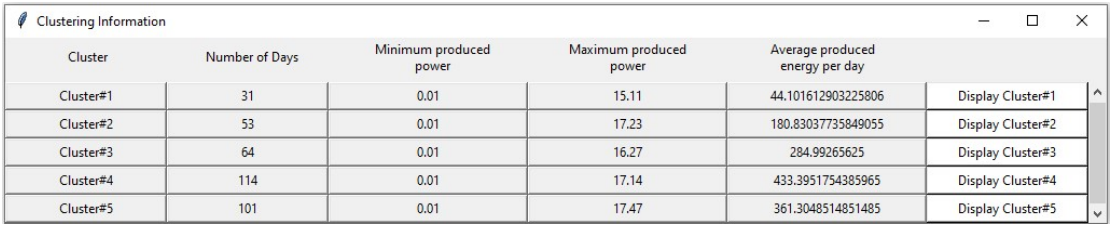
In order to confirm the changes and permanent delete the day from the data the user should press the ***"Confirm Changes"***. The following message will appear.



Even if the user does not want to delete any days from the dataset, he should press the button ***"Confirm Changes".*** After that, all the windows with the cluster information close automatically and the user returns to the main ***"Clustering"*** window. If the user wants to discard the changes he made, he should just close the window and repeat the previous process. If he decides to save the changes and update the data, it should be noticed that he can no longer edit the observations or delete outliers and the button ***"Edit Clusters"*** is deactivated.

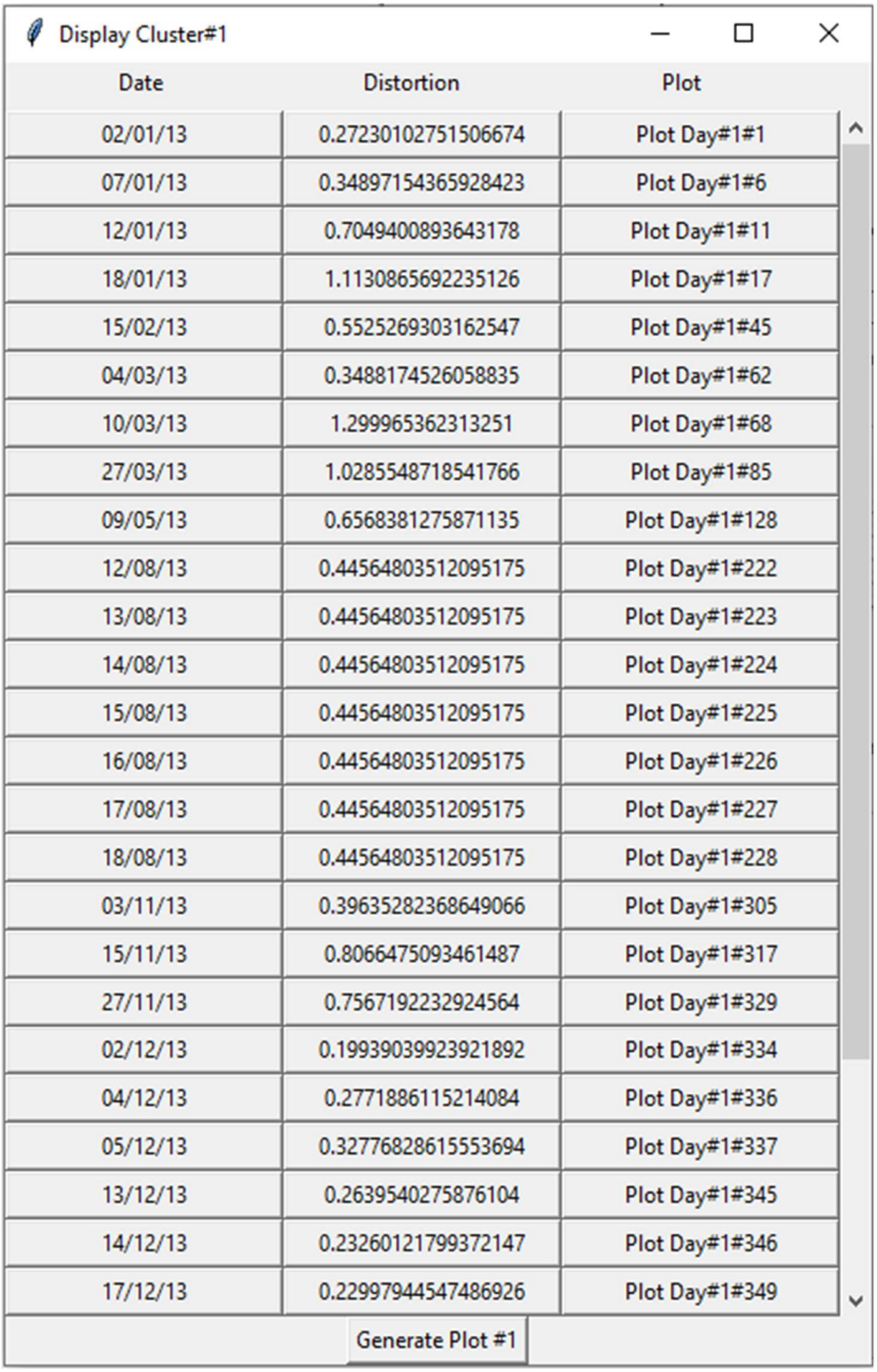


However, he can view the results of the clustering, after the delete of the invalid days, by pressing the button ***"Display cluster details"***. A new ***"Clustering information"*** window opens the DBSCAN clustering is not implemented and there no warnings about invalid data or singleton clusters. Instead, all the ***"Display"*** buttons are white.



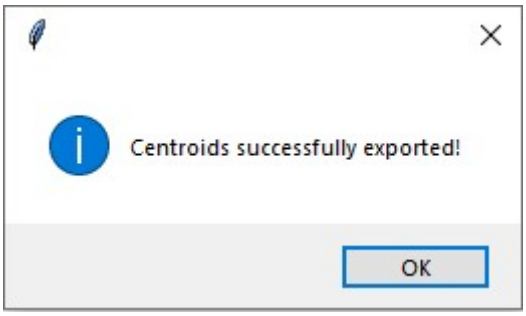
Cluster	Number of Days	Minimum produced power	Maximum produced power	Average produced energy per day	
Cluster#1	31	0.01	15.11	44.101612903225806	Display Cluster#1
Cluster#2	53	0.01	17.23	180.83037735849055	Display Cluster#2
Cluster#3	64	0.01	16.27	284.99265625	Display Cluster#3
Cluster#4	114	0.01	17.14	433.3951754385965	Display Cluster#4
Cluster#5	101	0.01	17.47	361.3048514851485	Display Cluster#5

If the user presses the ***"Display Cluster#1"*** button the following window will appear. This time, the user can plot all the days within the cluster and plot each day in respect to the cluster's centroid but he cannot delete the days.



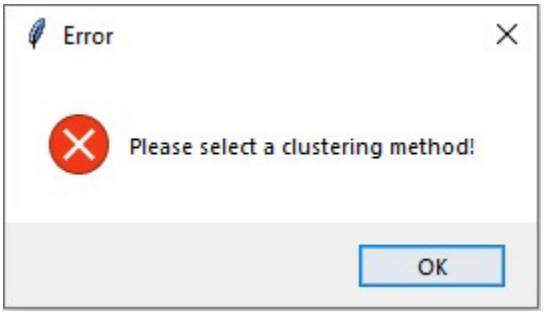
Also, the user can freely select another clustering method and number of clusters and just view the results of the clustering by pressing the ***"Display clusters details"***.

Finally, the user by pressing the ***"Save clusters centroids"*** he is able to export the centroids of the clusters to a ".xlsx" file. If the fille is created the following message will appear:

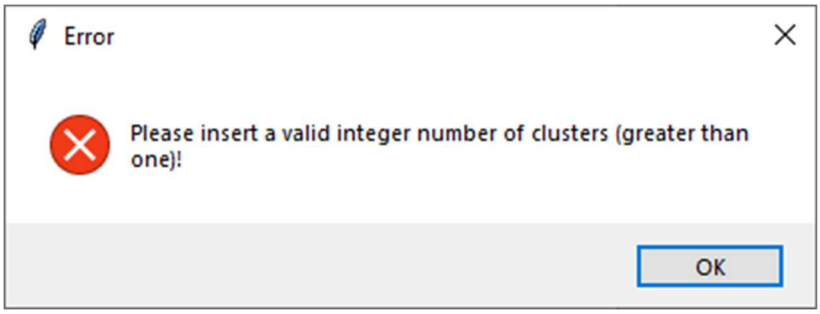


If the user does not complete all the necessary fields before he presses the ***"Edit clusters"*** button the following errors will appear:

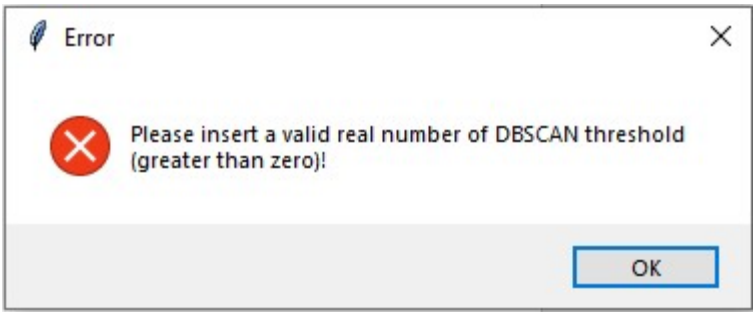
1. If the user does not select the clustering method.



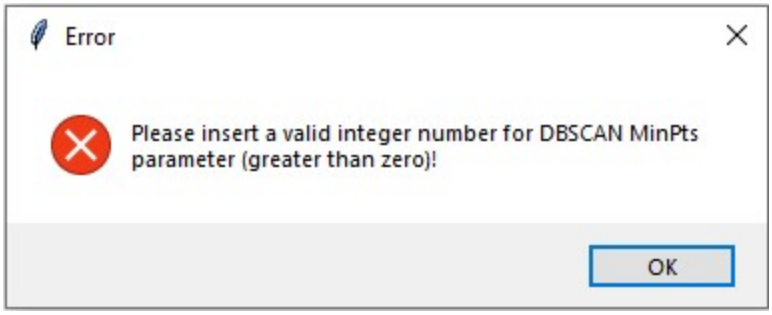
2. If the user does not select a valid number of clusters.



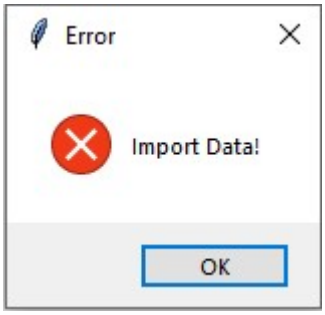
3. If the user does not select a valid number of DBSCAN epsilon parameter.



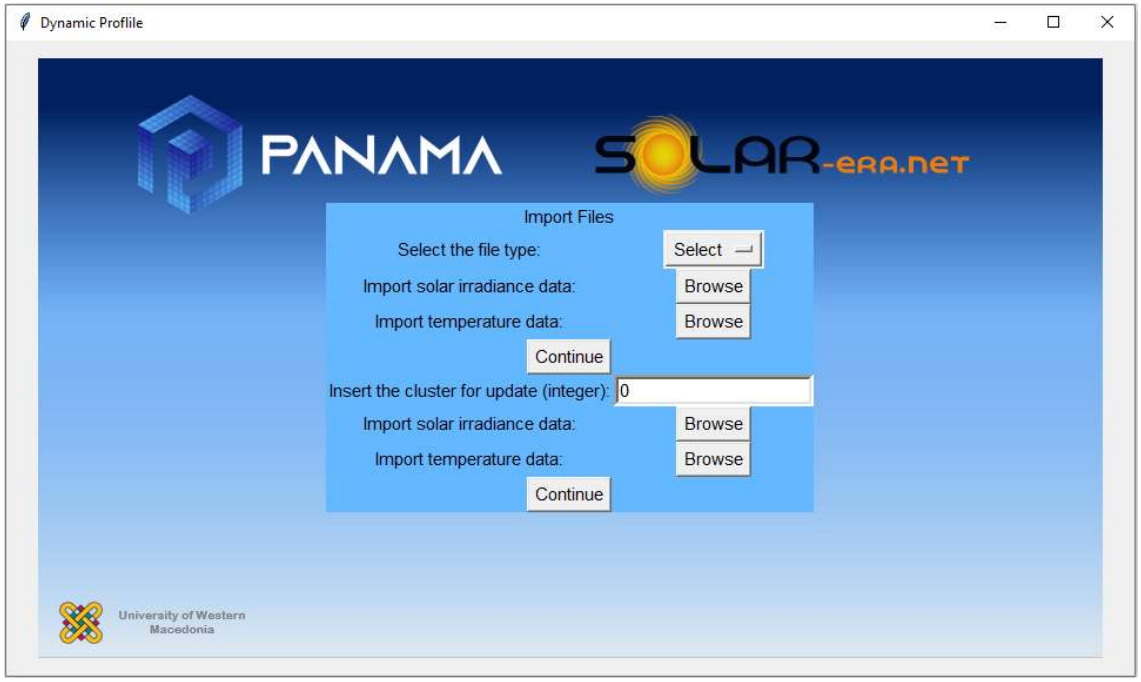
4. If the user does not select a valid number of DBSCAN *MinPts* parameter.



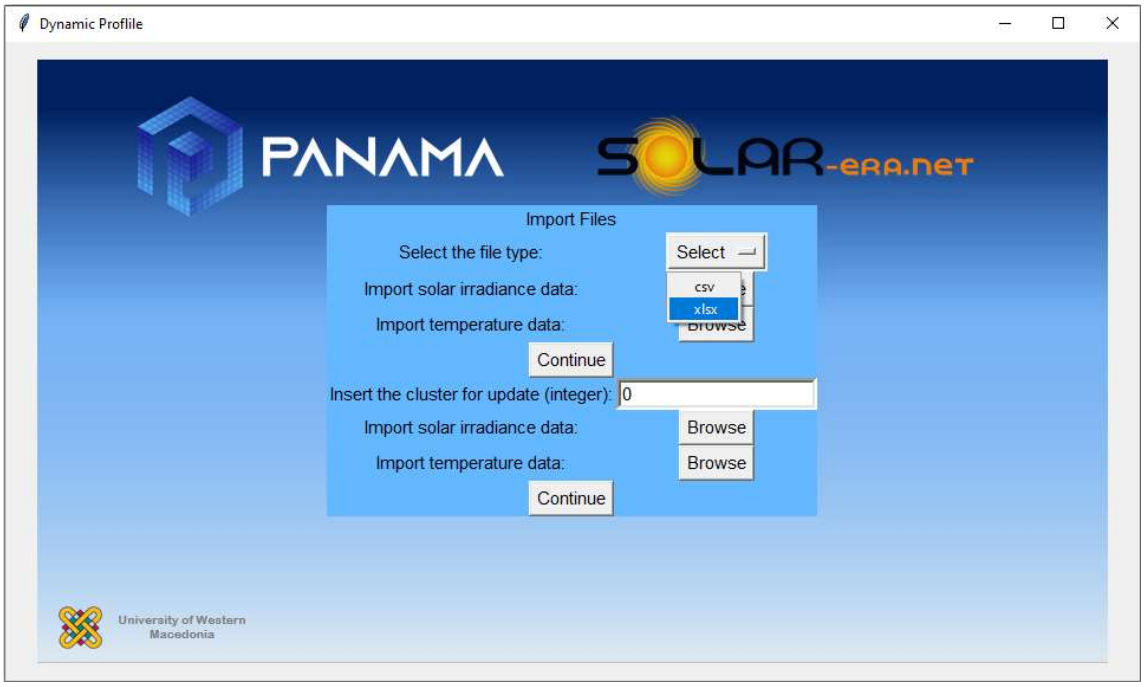
5. If the user does not select a file at the tab “Data”.



At the ***“Clustering”*** tab, after the editing of the clusters, a new ***“Dynamic profiles”*** button is appeared. If the user presses it, a new window will open. The dynamic profiles are based on the Linear Regression method and the user should have the solar irradiance and the temperature (module’s or ambient) that refer to the power data which were imported at the tab ***“Data”*** of the main ***“Clustering”*** window, with the same recording frequency.

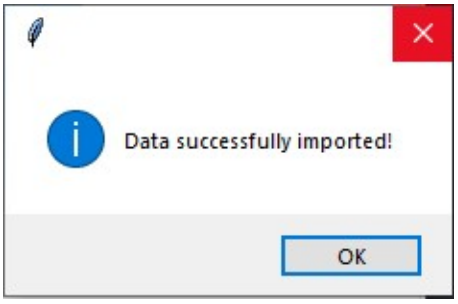


At first the user is called to select a file type at the field ***“Select the file type”***, which can be either “.csv” or “.xlsx”.

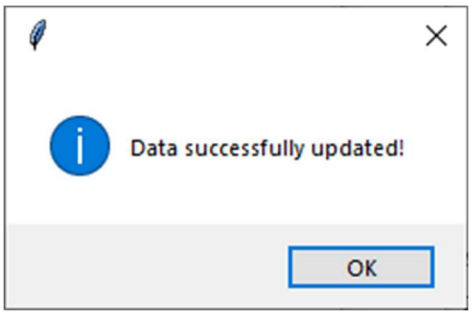


Afterwards, the user should select from the computer and import the first two files of the solar irradiance and the temperature by pressing the two ***"Browse"*** buttons respectively. The two first files refer to the generation data, which have been utilized for the static clustering process. Each time the user selects a file a browse window is appeared.

After the selection of each file the following message will appear in order to inform the user that the files are imported successfully.

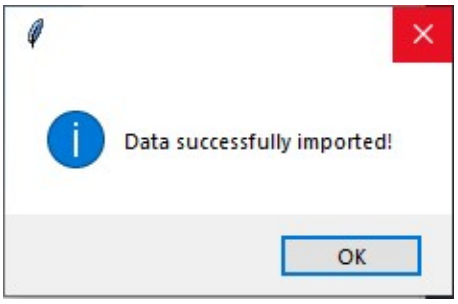


After the two first files selection the user it is necessary to press the ***"Continue"*** button in order to update the solar irradiance and temperature data according to the removal of the invalid days, which was implemented previously by the user. When the user presses the ***"Continue"*** button the following message appears.

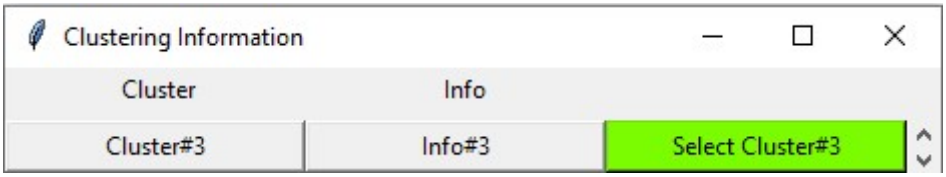


Accordingly, the user should select the number of clusters in order to apply the intra cluster linear regression method. The method utilizes the PV power, solar irradiance and temperature data of the days that belong to the cluster. It has to be pointed out that if the user, for example, wants to create a dynamic profile of power production at the 14th of June of 2014 he must check from the static clustering process, which was implemented previously, the cluster to which the 14th of June of 2013 belongs.

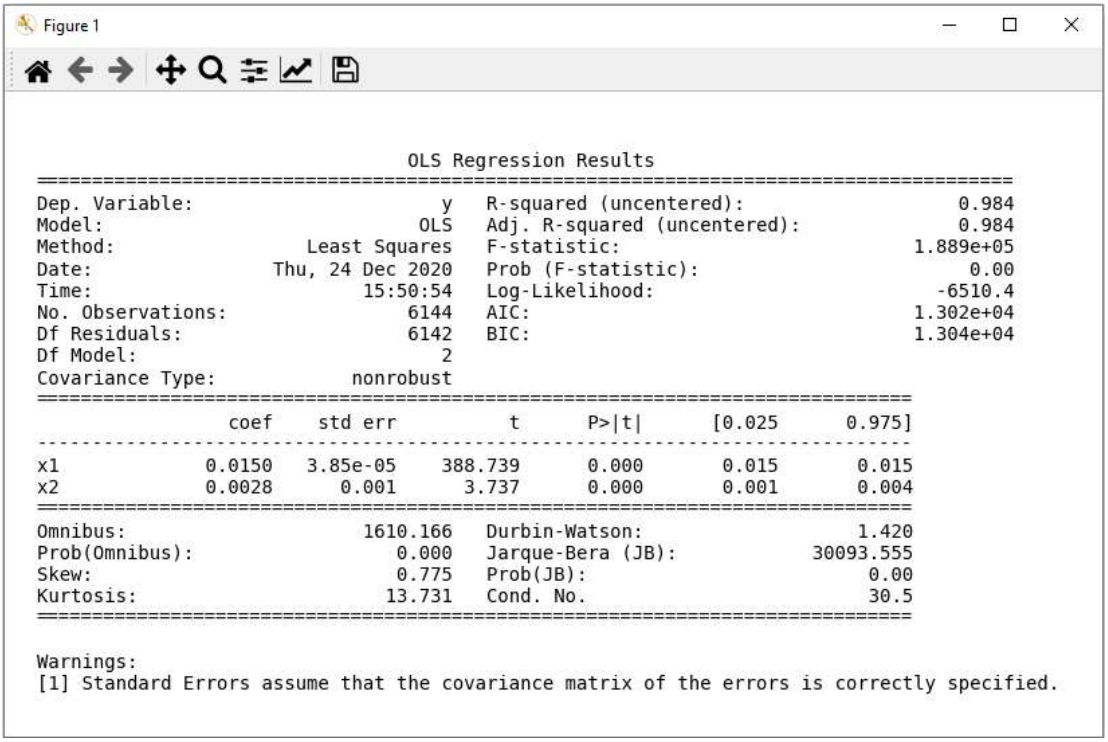
After that, the user should import the data of the solar irradiance and temperature of the 14th of June of 2014. After the selection of each file the following message appears in order to inform the user that the data are imported successfully.



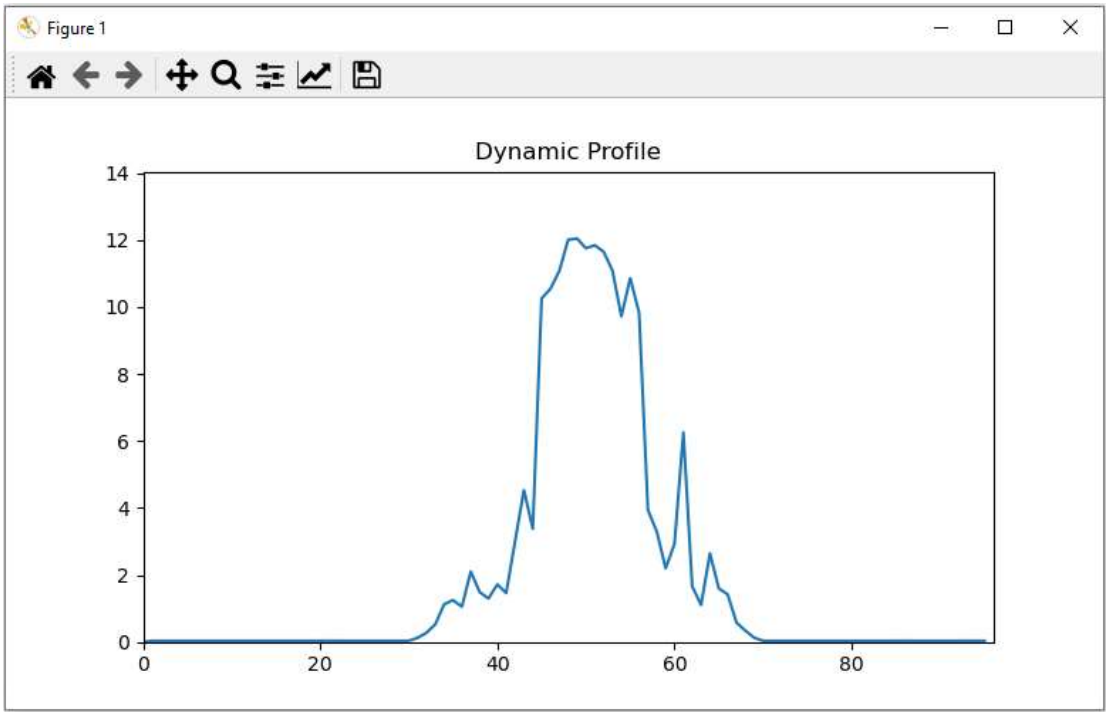
When all the four files are imported the user should press the ***"Continue"*** button in order to continue with the ***"Dynamic Profile"*** process and a new window will open. The first column contains the number of the selected cluster.



If the user presses the 'Info#3' button the detailed statistical information about the results of the linear regression.

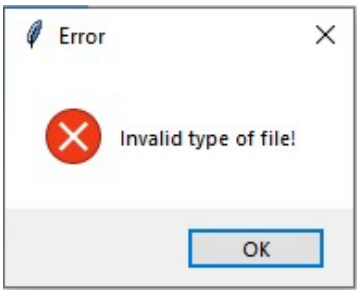


Finally, the ***"Select Cluster#3"*** button provides the power generation curve of the 14th June of 2014.

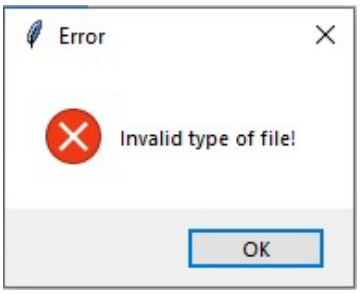


The GUI at the dynamic profile process provides the following error messages.

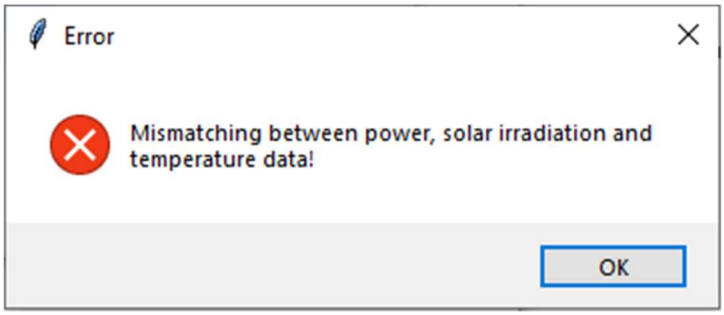
1. If the user does not select a valid type file from the browse window.



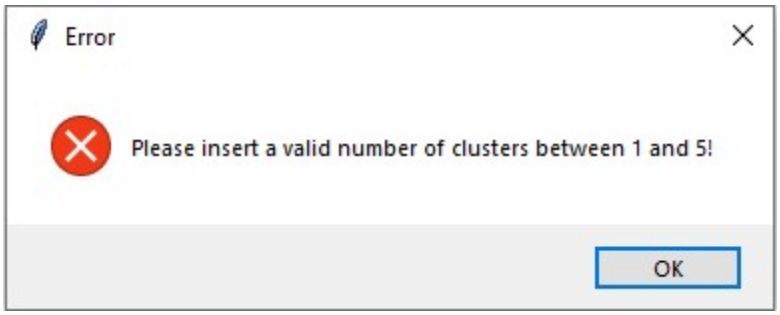
2. If the user does not import all the necessary data.



3. If the user selects solar irradiance and temperature data that do not refer to the power data which were utilized at the static clustering, or do not have the same recording frequency.



4. If the user selects an invalid number of clusters.



REFERENCES

[1] Wang W., Zhanga Y., 2007. On fuzzy cluster validity indices. Fuzzy Sets Syst. 158, 2095-2117.

[2] Xu R., Wunsch D., 2006. Clustering. John Wiley & Sons, Inc, New Jersey.

[3] Chicco G., 2012. Overview and performance assessment of the clustering methods for electrical load pattern grouping. Energy 42, 68-80.

[4] Weisberg S., 2005. Applied linear regression. John Wiley & Sons, Inc, New Jersey.

[5] Python™ programming language: https://www.python.org/