

# User manual - GisSOM

Authors: **Sakari Hautala, Jaakko Madetoja, Johanna Torppa**

Institution: **Geological Survey of Finland**

Date: **22.6.2021**

## TABLE OF CONTENTS

1	Introduction .....	3
1.1	Self-organizing maps and k-means.....	3
2	Installation.....	4
2.1	Installation requirements .....	4
3	Using the software .....	4
3.1	Load the data .....	5
3.2	Data preprocessing.....	6
3.3	Choose SOM and k-means parameters .....	9
3.4	Results .....	12
4	References.....	21

## LIST OF FIGURES

Figure 1.	Selecting the data format.....	6
Figure 2.	Data preprocessing.....	7
Figure 3.	Defining limiting values for winsorizing .....	7
Figure 4.	Choosing parameters for SOM and k-means .....	8
Figure 5.	Structure of the result folders. Blue boxes represent folders and orange boxes represent files. .....	12
Figure 6.	SOM space results .....	13
Figure 7.	Geospace results .....	14
Figure 8.	Clustering results.....	15
Figure 10.	Boxplot results.....	16
Figure 11.	Scatterplot results .....	17
Figure 12.	Interactive plot .....	18
Figure 13.	Selected cluster is drawn on the right side image .....	18
Figure 14.	Interactive plot is not showing .....	20
Figure 15.	Interactive plot in web browser .....	20

# 1 INTRODUCTION

The purpose of this document is to explain how to install, use and read the results of the *GisSOM* software developed in the European Union funded H2020 project NEXT. *GisSOM* is developed for performing self-organizing maps and k-means clustering for multivariate data. The software provides means to do simple preprocessing to the input data and select variables for clustering. Result visualization is applicable also for spatial data, showing the results in the original geospace in addition to the SOM space. Results are also available as csv tables with mapping from SOM space to geospacer. The calculation and results use 32-bit floating point numbers, so the results have around 7 digits of significance. Detailed information on the software design and algorithms used is provided in D 4.11 Appendix 2.

## 1.1 Self-organizing maps and k-means

Self-organizing maps (SOM) is an unsupervised artificial neural network that arranges a set of  $n$ -dimensional vectors to a usually two dimensional SOM lattice (Kohonen, 2001). The usability of SOM comes from its topology preserving nature: similar data vectors are assigned to SOM cells that are close together.

Although SOM can be considered as a clustering method itself, the number of cells in a SOM is generally too large for practical data classification, for instance. To reduce the number of clusters, *GisSOM* applies k-means clustering to the SOM result. K-means is a very basic clustering method where each data point is assigned to the cluster that best represents the data point, without considering the relation between or similarity of different clusters.

SOM and k-means computations are carried out using the *somoclu* package (Wittek et al., 2017). After the initialization of the SOM neuron weights, the training of SOM utilizes competitive learning (Kohonen, 2001): For a given data point, the neuron with the smallest Euclidean distance is found; this neuron is called the best matching unit (BMU). The weights of the BMU and the neurons close to it are updated to be closer to the data point. The formula for updating the weights is

$$w(t + 1) = w(t) + \alpha(t)h(t)(x(t) - w(t))$$

where  $w(t + 1)$  is the new weight for a given neuron,  $w(t)$  is the old weight,  $\alpha(t)$  is monotonically decreasing coefficient (learning rate),  $h(t)$  is a neighborhood function, and  $x(t)$  is the input data value. The learning rate ensures that the area in which the weights are updated shrinks over time and the neighborhood function ensures that the update is smaller the farther away the neuron is from the BMU in SOM space.

The quality of SOM is usually measured using two quantities. The *topological error* describes how closely similar data vectors are located on SOM and the *quantization error* is a measure of the goodness of clustering of data vectors in each SOM cell. In *GisSOM*, computation of only quantization error is implemented so far.

After SOM calculation, k-means clustering can be applied to its neurons. K-means is a clustering method that tries to minimize variances within clusters. The algorithm is iterative; it assigns observations to the closest cluster centroid and recalculates the centroids. This is repeated until no updating happens.

The user provides the minimum and maximum number of clusters for which k-means clustering is tested. As the initial random assignment of clusters affects the results of the algorithm, k-means is run multiple times (user provides the number of initializations) for each number of clusters in the given range. The best clustering result for each number of clusters is saved, and the three best clustering results are shown in the user interface, and stored. The goodness of clustering is measured using the Davies-Bouldin index (Davies & Bouldin, 1979).

## 2 INSTALLATION

The software comes with an installer. Double-click the installer to start the installation wizard and install the software.

### 2.1 Installation requirements

GisSOM requires Windows operating system (7, 8 or 10).

## 3 USING THE SOFTWARE

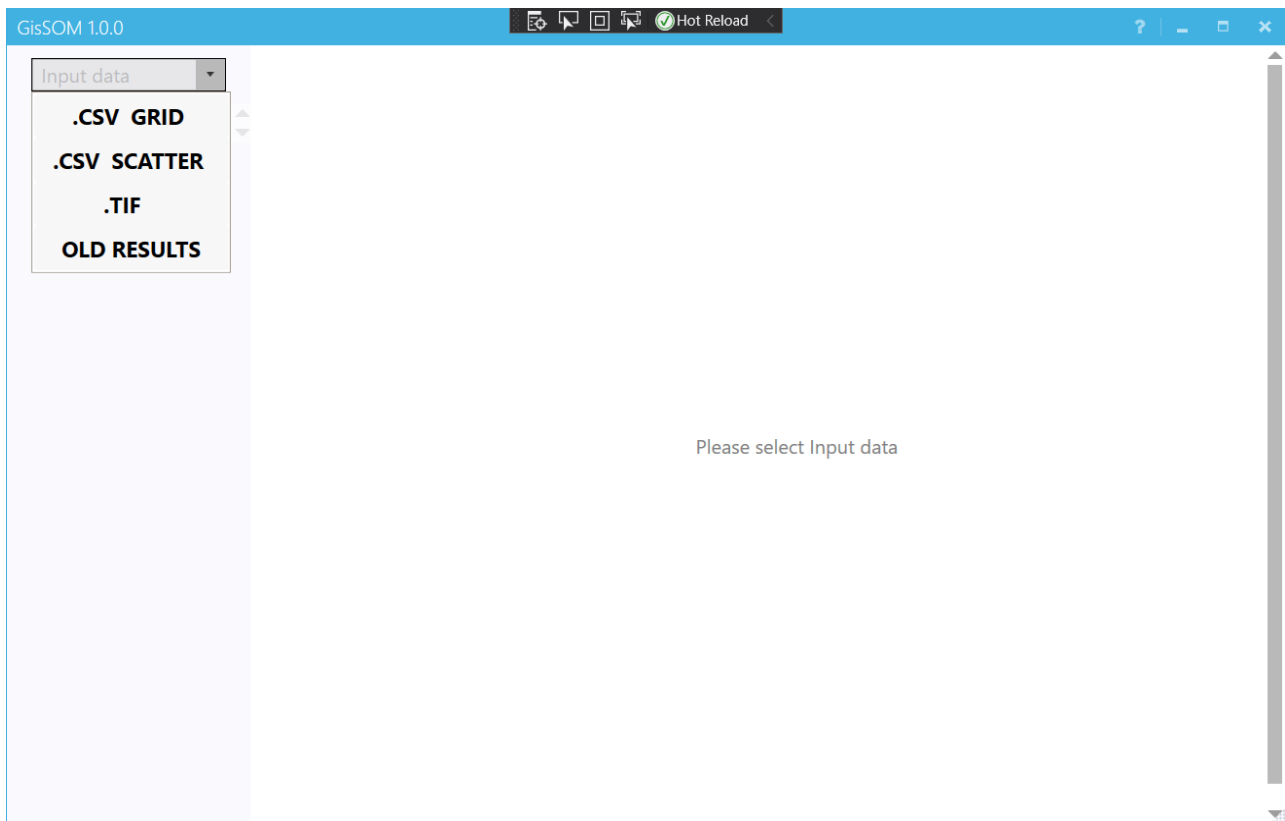
Open the software using the GisSOM icon that is created on your desktop, or by running executable SomUI.exe, which is in the root of the installation directory .../GisSOM. This will open a wizard-style window where you can load and study input data, perform simple transformations, provide SOM and k-means parameters, and study the results in SOM space, geospace, boxplots and scatterplots.

### 3.1 Load the data

In the first step of the wizard, you need to select the input data format (Figure 1) and locate the data file(s) from your computer. When you click an item in the data format dropdown menu, a file browser window opens where you can select your input data file(s). CSV format uses a single input file that contains all the input data parameters, but in the case of GeoTIFF files, multiple files can be given, each containing a single data parameter. When opening an CSV file, make sure the file is not open in Excel or any other spreadsheet, because then it can't be accessed.

Options are

- **CSV Grid:** A comma-delimited text file with comma (,) as the column separator and point (.) as the decimal separator. The file must have one header line containing the column names, and all columns have to have a non-empty header. Also, spaces (" "), percentage signs "%" and double quotation marks ("") are removed from headers. See also the example testdata file in .../GisSOM/TestData.
- **CSV Scatter:** Use this, when the data is not evenly spaced in a grid. If there are only individual missing values or missing values in the sides and center, CSV grid can be used, but if whole columns and rows are missing, CSV scatter is needed (Figure 2). The main drawback of the scatter plotting function is that it's slower than plotting the regular heatmap grid.
- **GeoTIFF:** A georeferenced raster data format. Note that when using raster data, the alignment and size of the pixels needs to be the same in each GeoTIFF file. Also, as the file names are used as headers for the data, spaces (" "), percentage signs "%" and double quotation marks ("") are removed.
- **Visualize existing SOM results:** Use this option, if you want to visualize existing SOM results from GisSOM software. Navigate to the output folder GisSom, and select one timestamped result folder. Make sure you haven't changed the file names or paths of the results.



**Figure 1.**      *Selecting the data format*



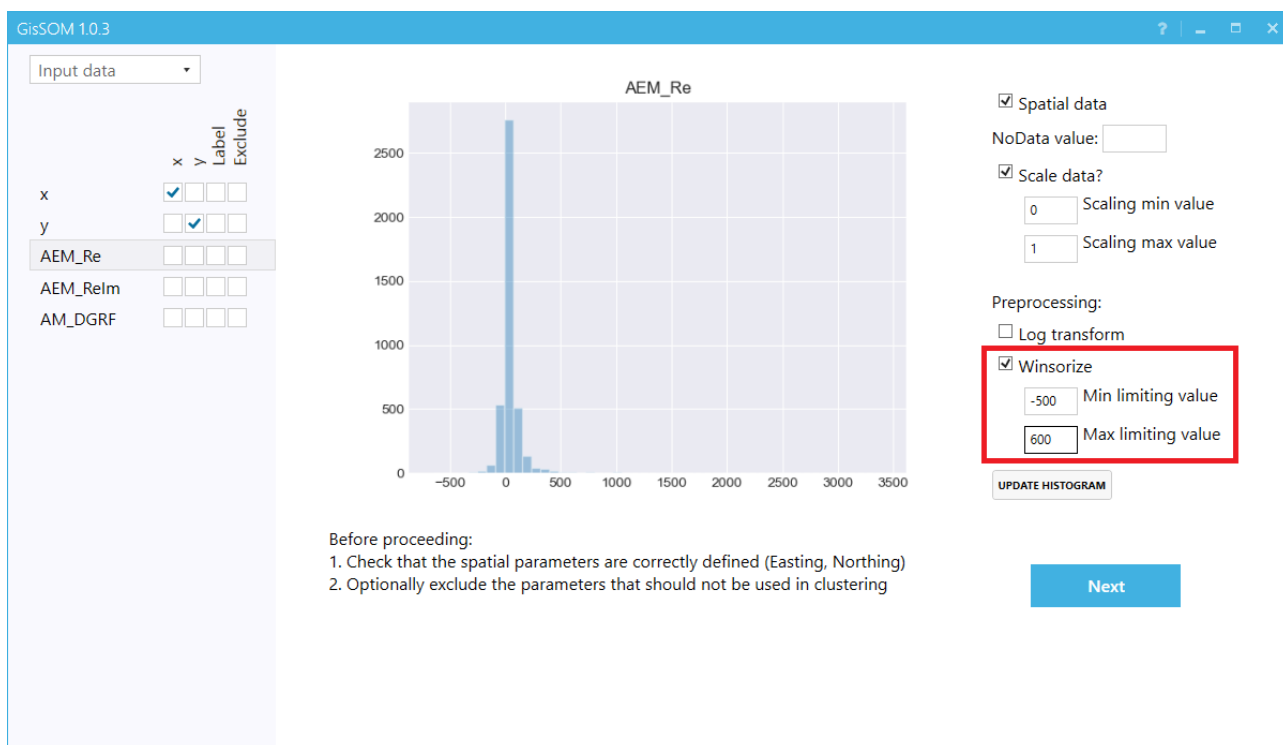
**Figure 2.** *Example of grid type and scattered data points. The data on the left can be used as grid data, as there are no missing columns or rows. The data on the right has a whole row missing, so CSV scatter needs to be used, otherwise the result plots will misrepresent the data.*

### 3.2 Data preprocessing

Once you have selected the input data file(s), the GisSOM window shows you all the relevant data parameters, including the possible spatial parameters (Figure 3). In the next step, you need to study the data before it can be used in SOM. You need to select the correct North and East coordinates in the case of CSV files, exclude any data you don't want to use, and possibly transform data. The number of datapoints represented in the histogram is capped at 5000 to speed up performance, if the number of data points in a variable exceed this number, a random sample of 5000 is taken for plotting.



**Figure 3.** Data preprocessing



**Figure 4.** Defining limiting values for winsorizing

The following steps are required:

- Select the North and East coordinates by checking the “x” and “y” checkboxes for the correct parameter from the left. This is not required for georeferenced raster data (GeoTIFF). If you are not using spatial data, untick the box “Spatial data?” on the right side. If you have labels in your data, i.e. a parameter that represents true values, mark them using “Label” checkbox.
- Exclude the data parameters that you don’t want to be used in SOM by choosing “Exclude” for the correct parameters on the left side. Note that geographic coordinates should be excluded, as you normally do not want to use them in the SOM as parameters.
- As it is important to scale the data parameters into a comparable range of values for SOM computation, each parameter is normalized by default to the range of [0,1]. If the user wants to weigh data parameters differently, she can provide parameter specific ranges. Scaling is done using a simple shift and scale function:

$$r = \frac{(r_{max} - r_{min}) * (x - x_{min})}{x_{max} - x_{min}} + r_{min}$$

where  $r$  is the rescaled value,  $x$  is the original value,  $x_{min}$  and  $x_{max}$  are the minimum and maximum original data values, and  $r_{min}$  and  $r_{max}$  are the rescaled values defined by the user. Before visualization of the results, the values are transformed back using inverse of the formula. If you do not wish to apply any kind of rescaling, uncheck the checkbox.

- Study the value distribution of the parameters that you want to use as attributes in SOM using the histogram in the middle. If a parameter histogram has long tails, you can apply logarithmic transformation (“Log transform”) or limit extreme values (“Winsorize”, Figure 4) to make the histogram closer to a normal distribution. Click “Update histogram” to see the changes in the histogram.
  - Logarithmic transform is carried out by first shifting all the parameter values to the positive range. Then, a natural logarithm is applied to the values and computation is performed using the transformed values. After the analysis, the values are inverted back using exponential function. This option can be used when the distribution of data is right-tailed and logarithmic transformation results in values that are closer to normal distribution.
  - Winsorizing means assigning a limiting value to parameter values below and above the given limiting value. This can be used, if it is known that it is enough to classify very large values as only “large” and very small values as only “small”, without consideration of how big or how small. Note that winsorizing will be applied to the values shown in the histogram; if you apply logarithmic transformation first, you need to enter the logarithmic values, i.e. the values that you see in the histogram.
- If the dataset contains a numeric value signifying null values, provide it in the NoData value textbox. Not doing so can lead to erroneous results, or to the functions failing altogether in case of infinite or extreme null values. GisSOM automatically screens for non-numeric values, so those do not have to be provided.
- Click “Next” to proceed to the next step



### 3.3 Choose SOM and k-means parameters

In the next step, you need to choose the parameters used in SOM and k-means clustering (Figure 5). The input parameters are separated into two sections, basic parameters and advanced parameters. By default only the basic parameters are visible on screen, but you can view the advanced parameters by clicking on the “Advanced Parameters” -button.

GisSOM 0.9.1

Som x 10

Som y 10

Map type toroid

Grid shape rectangular

Run k-means clustering ☒

Clusters 2 Min 25 Max

Initializations 5

Output folder C:\Users\jmadetoj\AppData\Local

Show Advanced parameters

Run Results

Epochs 10

Initial codebook Select file

Initialization random

Neighborhood function gaussian

Initial neighborhood 0

Final neighborhood 1

Training rate function linear

Initial training rate 0.1

Final training rate 0.01

Figure 5 Choosing parameters for SOM and k-means

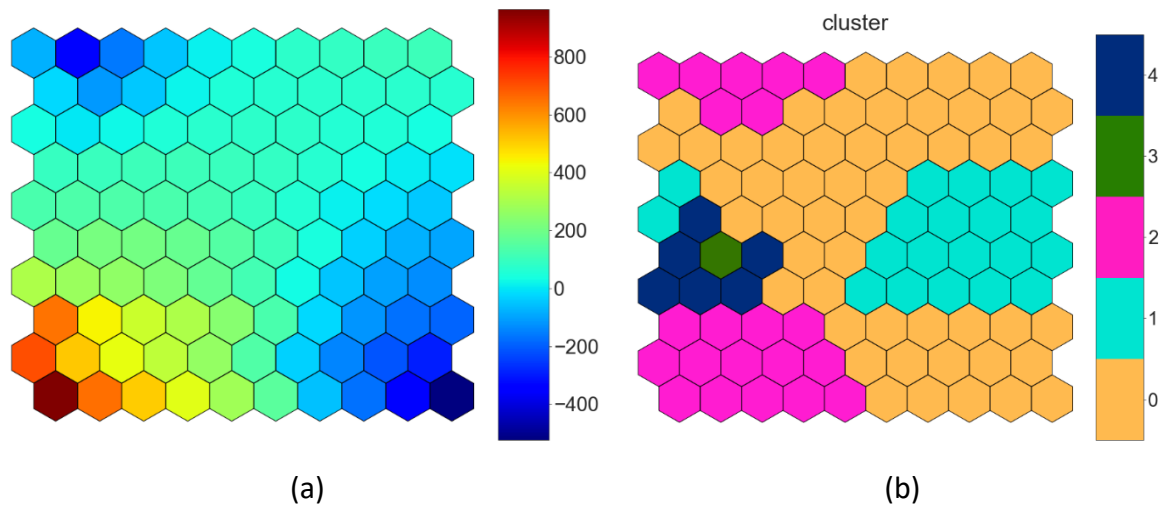
#### Basic Parameters for SOM

- Choose the size of the SOM using “Som x” and “Som y”. These refer to the number of SOM cells in horizontal and vertical direction. The default values are calculated using one rule of thumb for a square shaped map: the total number of cells is  $5 * \sqrt{\text{number of data points}}$  so both “Som x” and “Som y” are square root of that value. Large SOM size causes SOM cells to be closer to the input data and thus more accurate with the cost of computing time. Higher accuracy also means that quantization errors are smaller.
- Map type: topology of the map, accepted values are “toroid” or “sheet”. *Sheet* type SOM works in the same manner as SOM is visualized: the sides of the SOM map have an edge next to them and neighbors only on the other side. *Toroid* type SOM continues from one edge to the opposite, so that each SOM cell has the same number of neighbors. This is visualized in Figure 6.

		3		
1				2
		4		

**Figure 6. Neighbours in sheet and toroid type maps. For toroid 1 and 2 as well as 3 and 4 are neighbours; for sheet they are not.**

Using sheet type often causes large and small values to be clustered near an edge whereas toroid doesn't have this issue. Using toroid type often causes clusters to appear far away visually as they continue from the other side. These cases are visualized in Figure 7.



**Figure 7. Sheet type causing values to be piled near an edge (a) and toroid type causing a cluster to appear on the other side (b).**

- Grid shape: The shape of the grid that connects the nodes of the map. Accepted values: hexagonal or square. The software regards 4 closest SOM cells as neighbors in rectangular grid shape and 6 closest in hexagonal. The hexagonal grid starts from upper left corner and is not indented.

### Parameters for k-means

- You can choose to skip k-means clustering and run only SOM by removing the tick in “Run k-means clustering”.
- If you run k-means, you need to select the minimum and maximum number of clusters. The default for minimum is 2 and maximum is 25. This software applies k-means to the results of SOM using all values between these for the number of clusters and the most optimal number is chosen based on the smallest the Davies-Bouldin index.

- Choose the number of random “*Initializations*”. The default is 5. K-means utilizes random number generator in the algorithm and is sensitive to the initialization. Thus, this software runs k-means using different initializations and chooses the most optimal based on the smallest the Davies-Bouldin index.
- You can select a folder where all the results will be saved as “*Output folder*”.
- Click “*Run*” to run the software and after it, click “*Results*” to study the results.

### Advanced Parameters

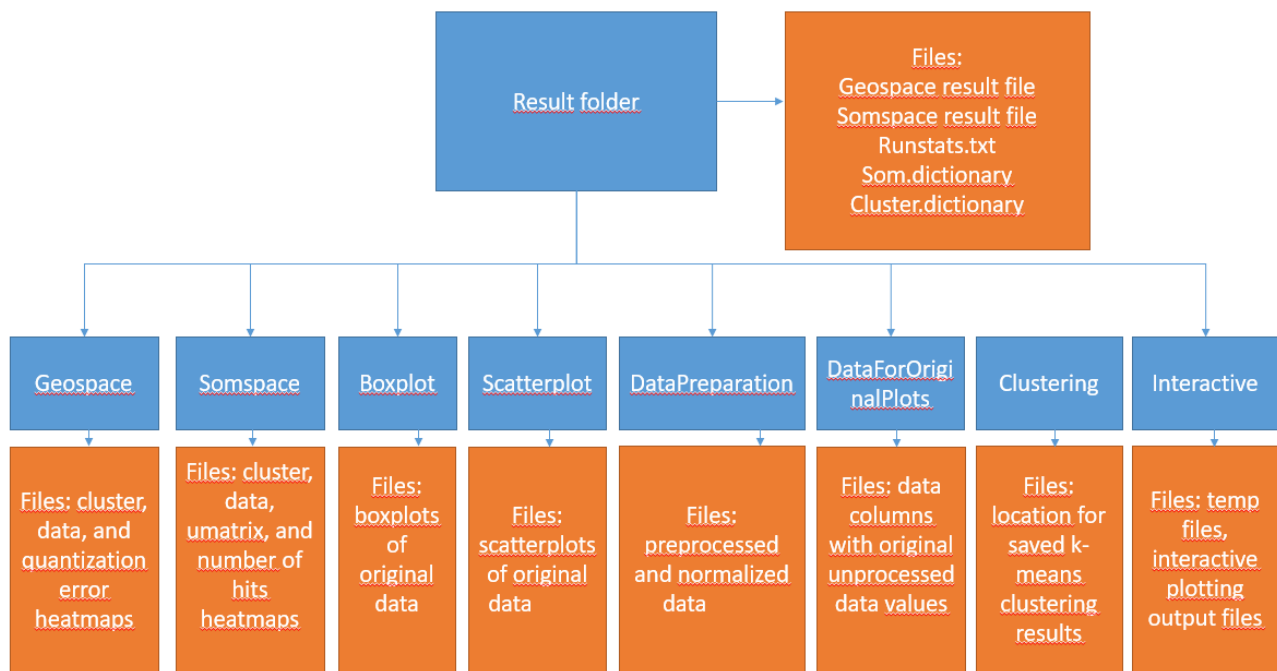
- “*Epochs*” is the number of times that the data set will be used when training the SOM. The default is 10. Small values result in faster computation, but possibly also in an inaccurate SOM. Larger values increase computation time, but might also improve the quality of the SOM. Usually the quality will not increase after certain amount of epochs.
- Initial codebook vectors: It is possible to provide a previously run SOM calculation codebook vectors as an initialization for the SOM. Codebook vectors are provided in the form of the “som.dictionary” file that the SOM calculation of GisSOM creates in the project output folder. The “som.dictionary” file is located in the time stamped result folder, under the main “GisSom” output directory. “GisSom” output folder is created according to the “Output folder” option in the user interface; the default is generally “C:/Users/YourUserName/AppData/Local/Temp/GisSom/”. The initial codebook vector matrix must have the same number of elements as the input dataset and have the same “Som x” and “Som y” parameters. If initial codebooks are provided, the value provided for the “initialization” parameter will be skipped altogether.
- Initialization: Initialization of the codebook vectors. Accepted values: “random” or “pca”. If you provide initial codebook vectors, this value will be skipped. “random” initialization of the codebook vectors picks random data points from the original data set and uses their attributes as the initial weights in SOM. “pca” initializes the weights from the first two eigenvectors of the correlation matrix. “pca” method also includes a random component, so the resulting maps might show small variation.
- Neighbourhood function, initial neighbourhood, final neighbourhood: The initial neighbourhood is the initial radius on the map where the update happens around a best matching unit, and the final training rate is the radius where the update happens in the final epoch. For the initial neighbourhood, the default value of 0 will trigger a value of  $\min(n\_columns, n\_rows)/2$ . Type of possible neighbourhood functions include “gaussian” and “bubble”. “gaussian” function utilizes a decreasing Gaussian function from the away from the BMU when the weights of SOM are updated. The software utilizes a cut off that ensures that weights beyond the training radius are not updated. “bubble” function is a simple constant function resulting in all neurons around the BMU getting the same proportional update.
- Training rate function, initial training rate, final training rate: Initial training rate is the training rate in the first epoch, and final training rate is the training rate for the last epoch.

The training rate function determines the cooling strategy between the initial and final learning rate, and possible values are “linear” or “exponential”.

### 3.4 Results

In the last step, you can see the results. These are divided between “*Somspace results*”, “*Geospace results*”, “*K-means clustering*”, “*Boxplots*”, “*Scatterplots*” and “*Interactive*”. You can access these using the buttons on top of the results window.

Results from different runs will follow the format Result\_timestamp where the timestamp is in the format “year-month-day, hour-minute-second” (for example “Results\_20200831125525”). This result folder includes the SOM space and geospace results (result\_som.txt and result\_geo.txt) and information on the run (RunStats.txt) as text files and dictionary files (som.dictionary, cluster.dictionary). Subfolders contain figures and/or data for geospace and SOM space results, boxplots, scatterplots, data preparation, original data, and the interactive plot. The folder structure is visualized in **Figure 8**.



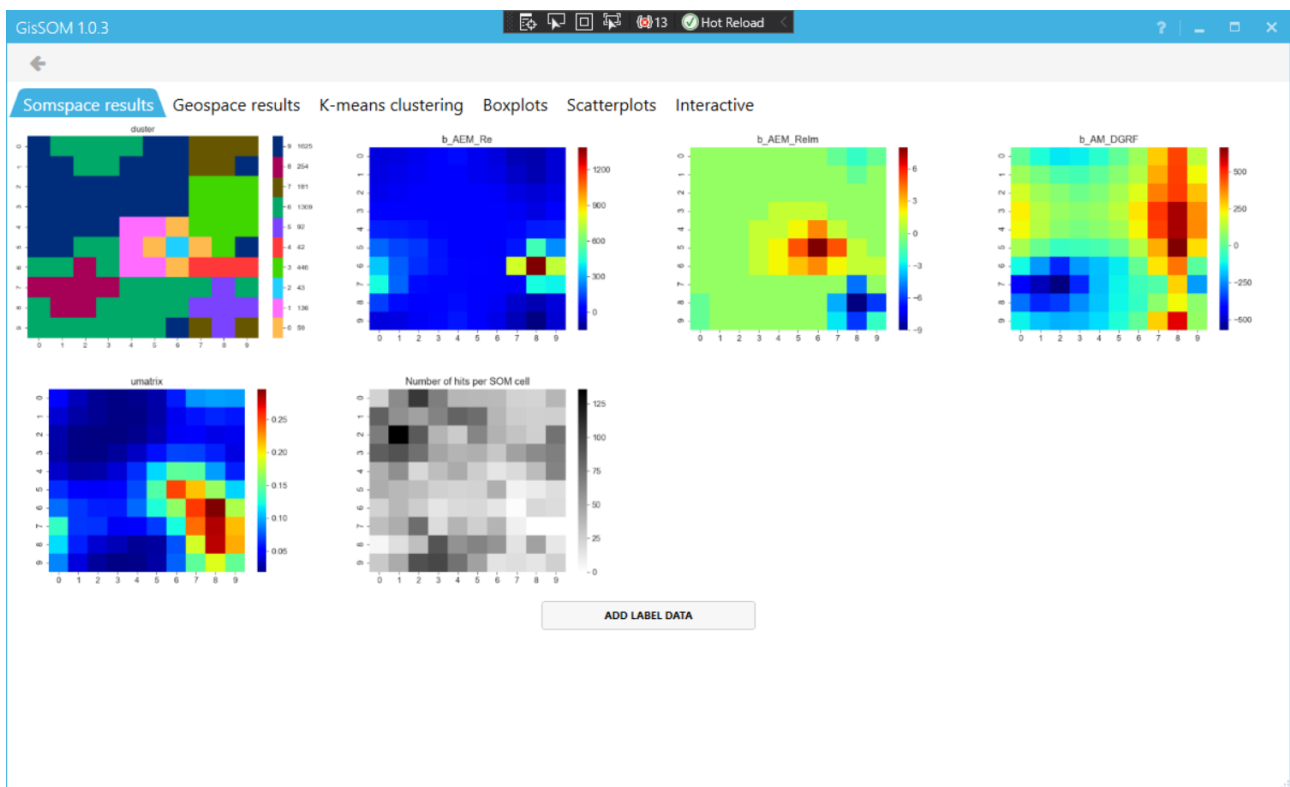
**Figure 8.** Structure of the result folders. Blue boxes represent folders and orange boxes represent files.

## Results in SOM space

These images (Figure 9) show the resulting SOM, color coded using various parameters:

1. value of each codebook vector element (corresponding to data parameters)
2. u-matrix (the magnitude of the difference of the codebook vectors in neighboring SOM cells)
3. K-means cluster. The left-most numbers on the color bar represent the number of the cluster, and the numbers on the right represent number of data points per cluster.
4. number of data points clustered in each SOM cell

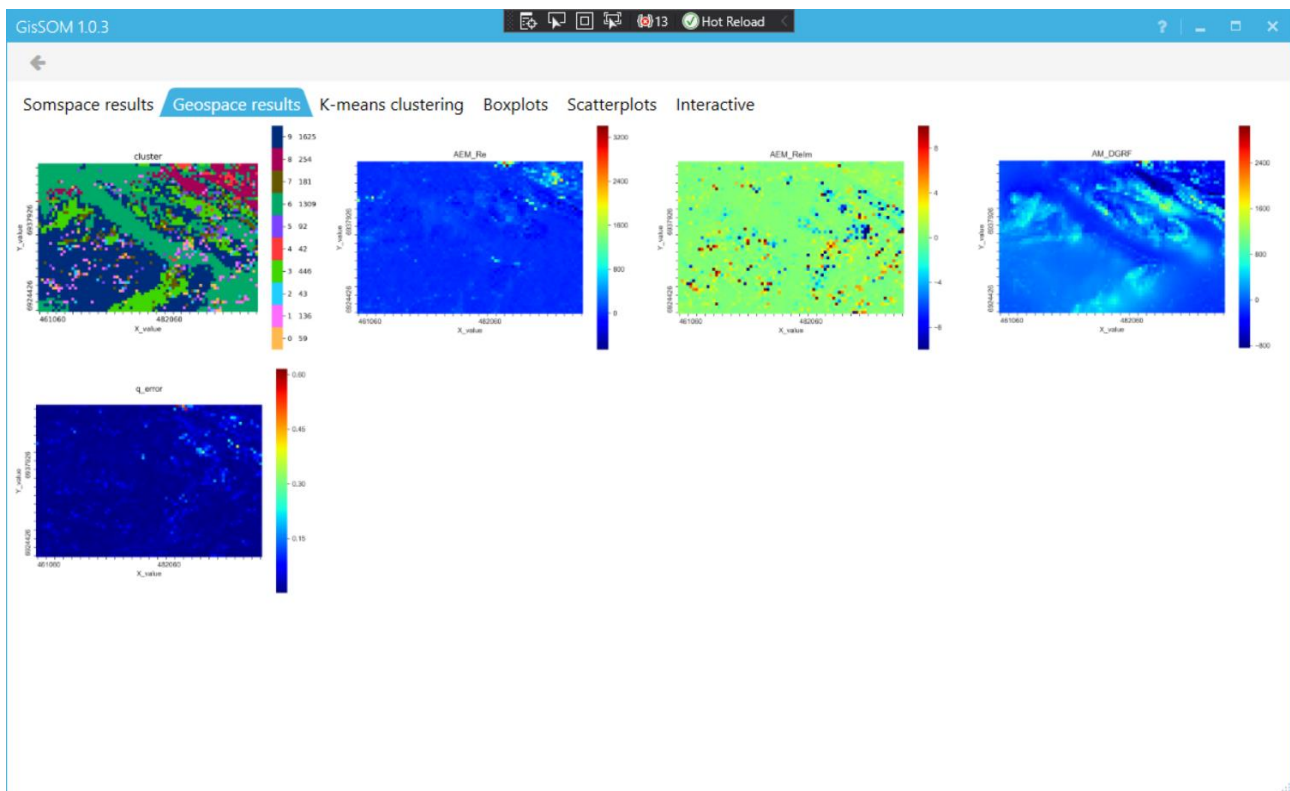
The “Add label data” button can be used to insert new datapoints with labels, and the data will be visualized in a new plot. The data must be provided in .CSV format, and the number and order of columns must match the columns of the somspace results file (result\_som.txt or result\_som.csv).



**Figure 9.** SOM space results

## Results in geospace

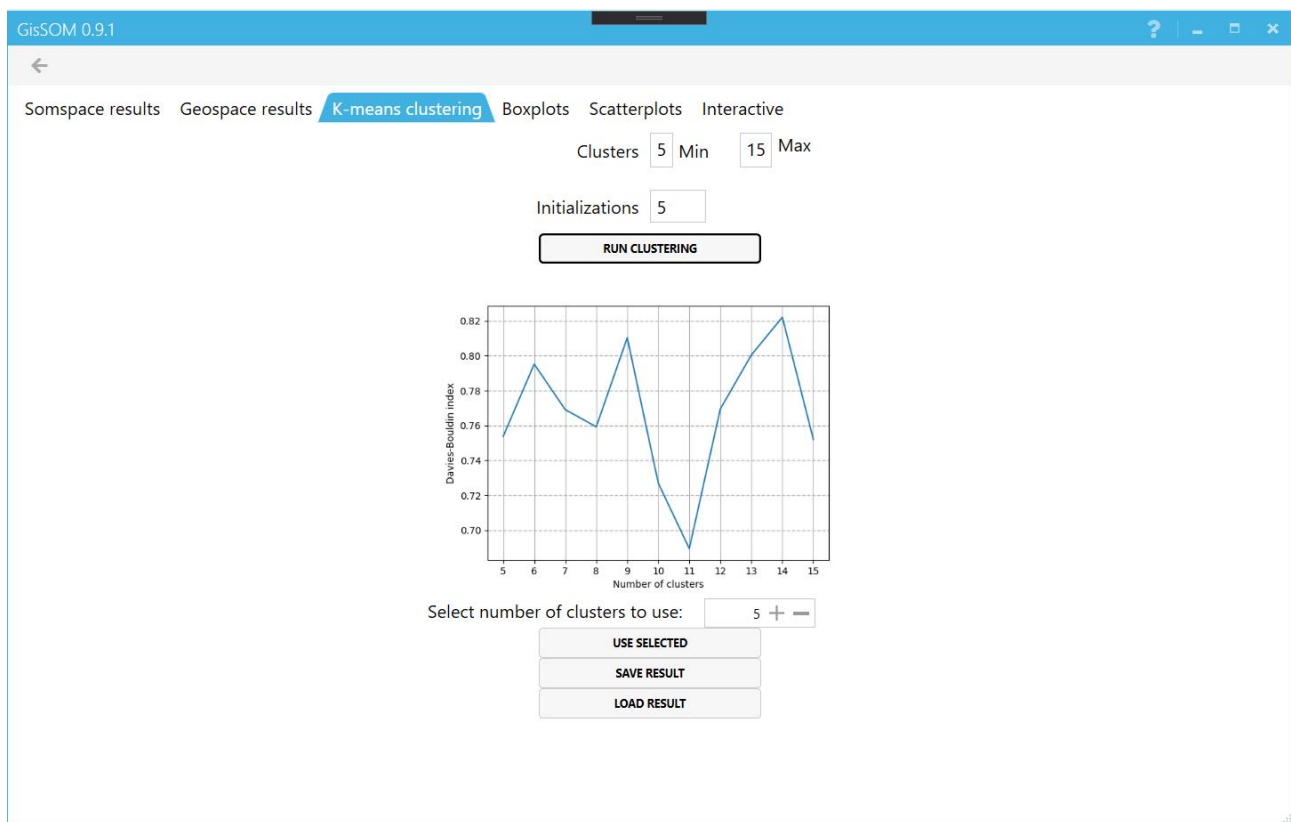
These images show the k-means clustering results, original attributes and quantization errors in geographical space (Figure 10). Quantization error is the difference between the original data parameter values and the SOM codebook vector elements of the cell where the data point has been projected to. High quantization error values show outliers in the data.



**Figure 10. Geospace results**

### *K-means clustering*

If you are not satisfied with the current clustering, or want to explore different clustering variants, it can be done in the K-means clustering tab (Figure 11). You can run the clustering again with different minimum and maximum number of clusters or initializations. The plot shows the Davies-Bouldin index for each number of clusters. Use “Select number of clusters to use” and click the “Use selected” button to get results for a different number of clusters. The other tabs will be redrawn accordingly.



**Figure 11.** Clustering results.

### Boxplots

These images show different data parameters of the k-means clustering results as boxplots (Figure 12). Color coding is the same as in the geospace and somspace images.



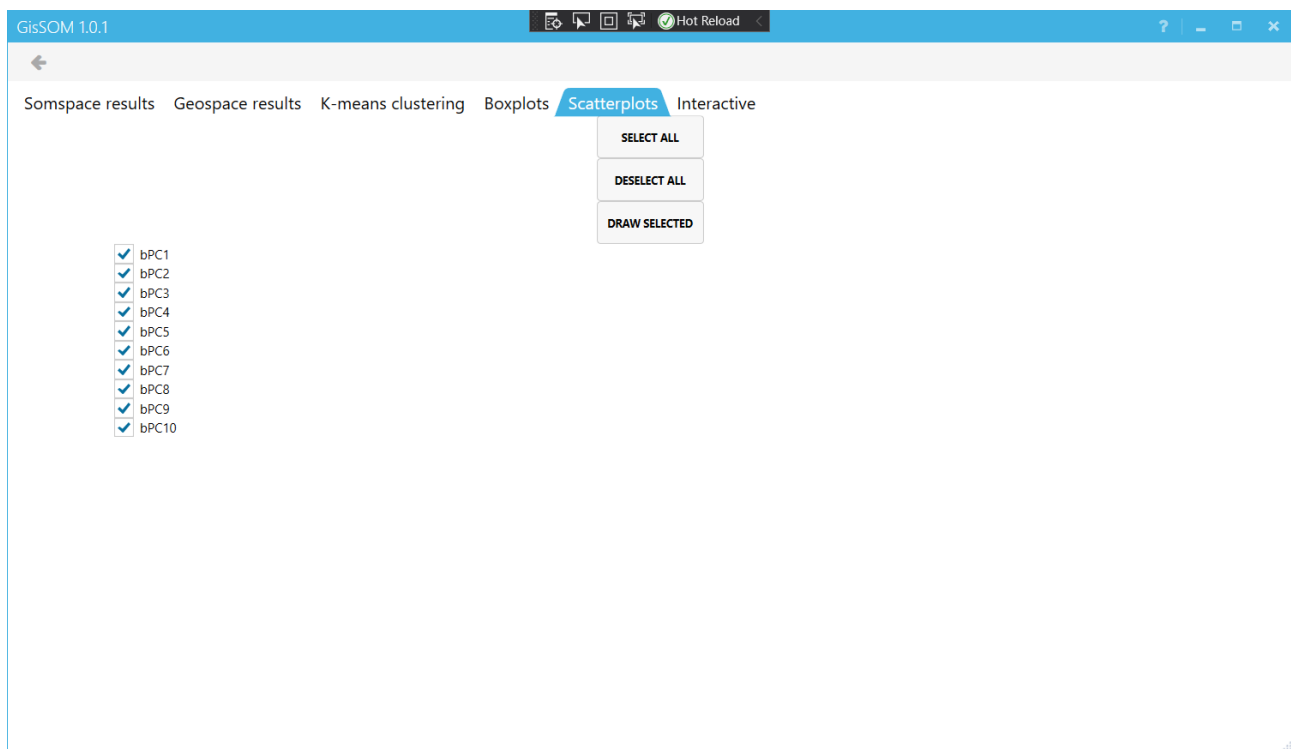
**Figure 12. Boxplot results**

The window includes one image for each data parameter. In an image, there is one boxplot for each cluster. The boxplot describes the distribution of original data values: the line in the middle of the box is the median value and the borders of the box are the first (25 %) and third (75 %) quartiles. The lines extend from the box borders to reach the minimum and maximum values, but no more than 1.5 times the size of the box. If there are any points outside the box and lines, they are visualized using discrete points.

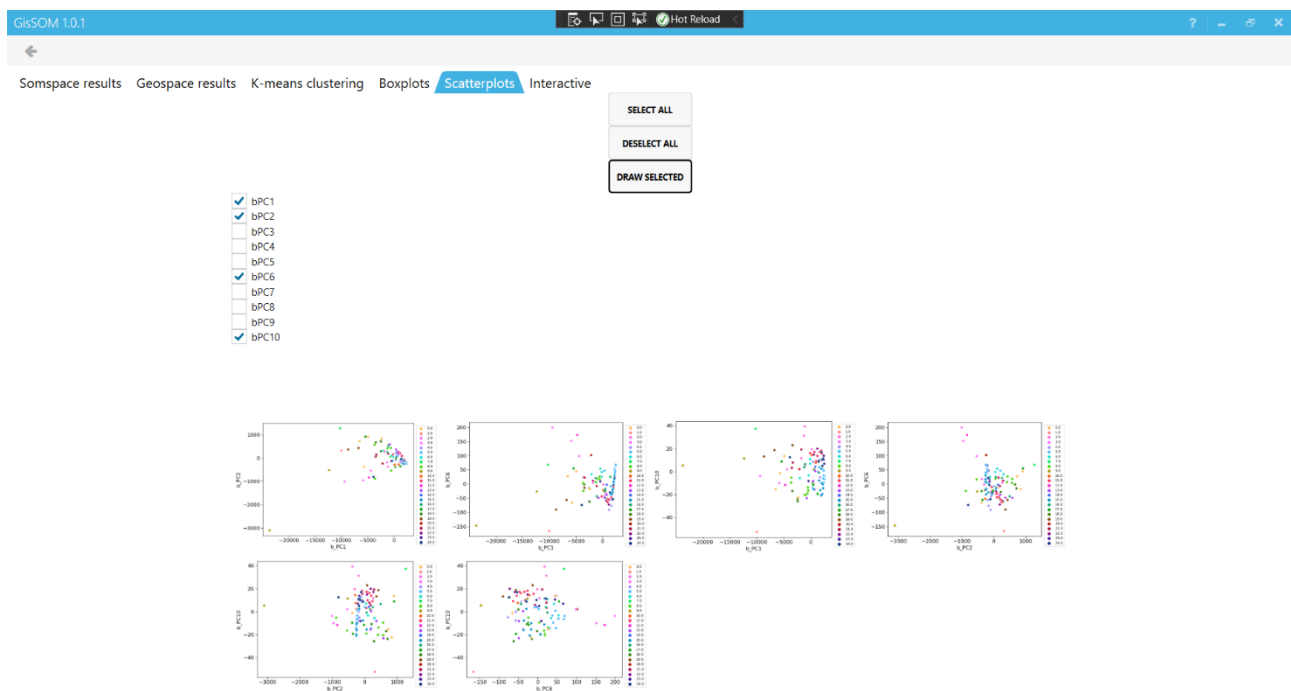
### Scatterplots

Scatterplots are not plotted by default, as their number increases exponentially with the number of columns in the input data and depending on the dataset this can lead to a very large number of plots. Select the variables you want to plot and click “Draw Selected” to draw the scatterplot pairs you want. Each image shows one pair of data parameters as a scatterplot (Figure 13). Each point in the scatterplot represents the original data point in x,y-coordinates with one parameter as x and another as y. The dots are colored based on the cluster that the data point belongs to.





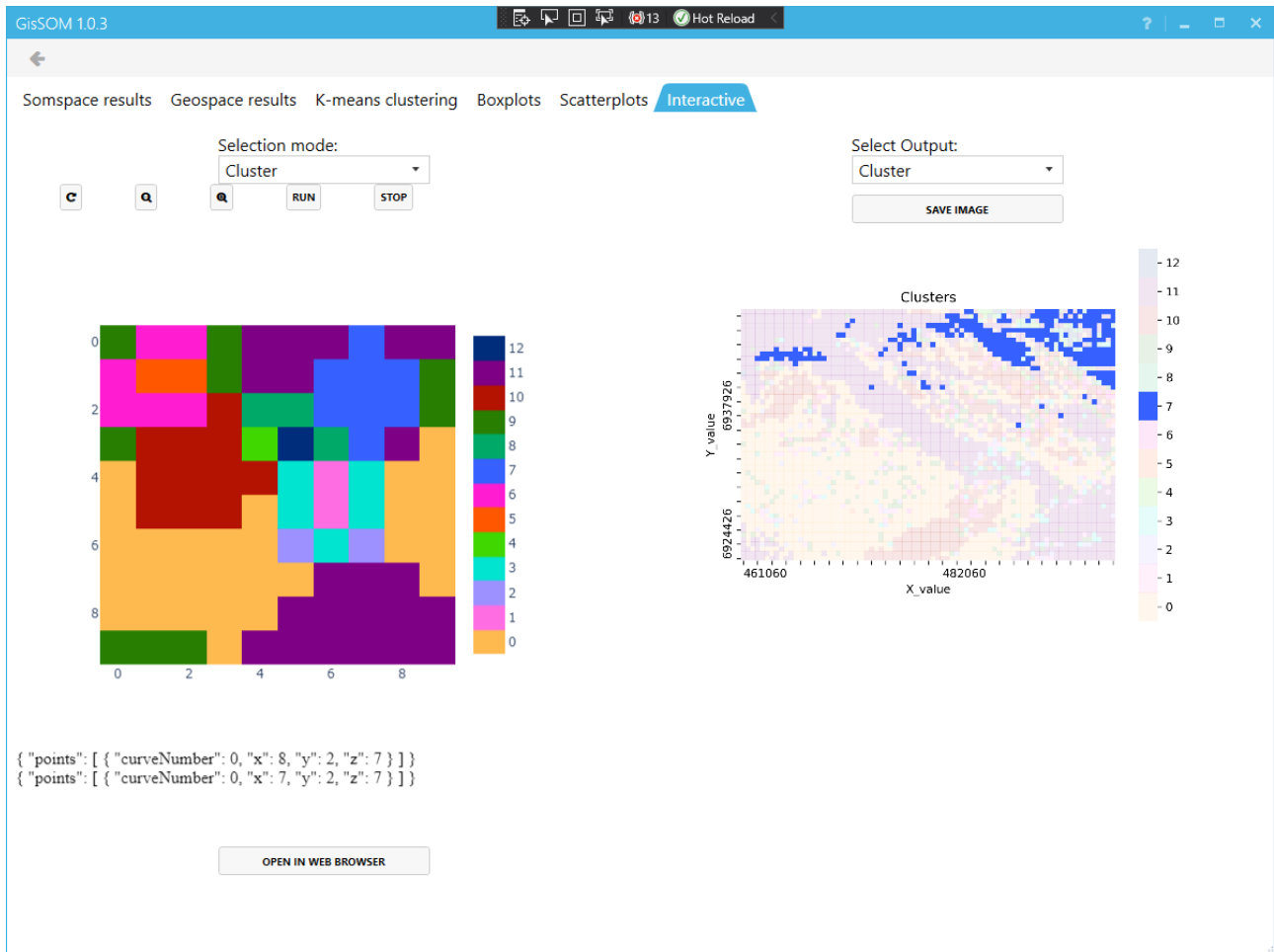
**Figure 13.** Scatterplot page after SOM calculation. No scatterplots have been drawn.



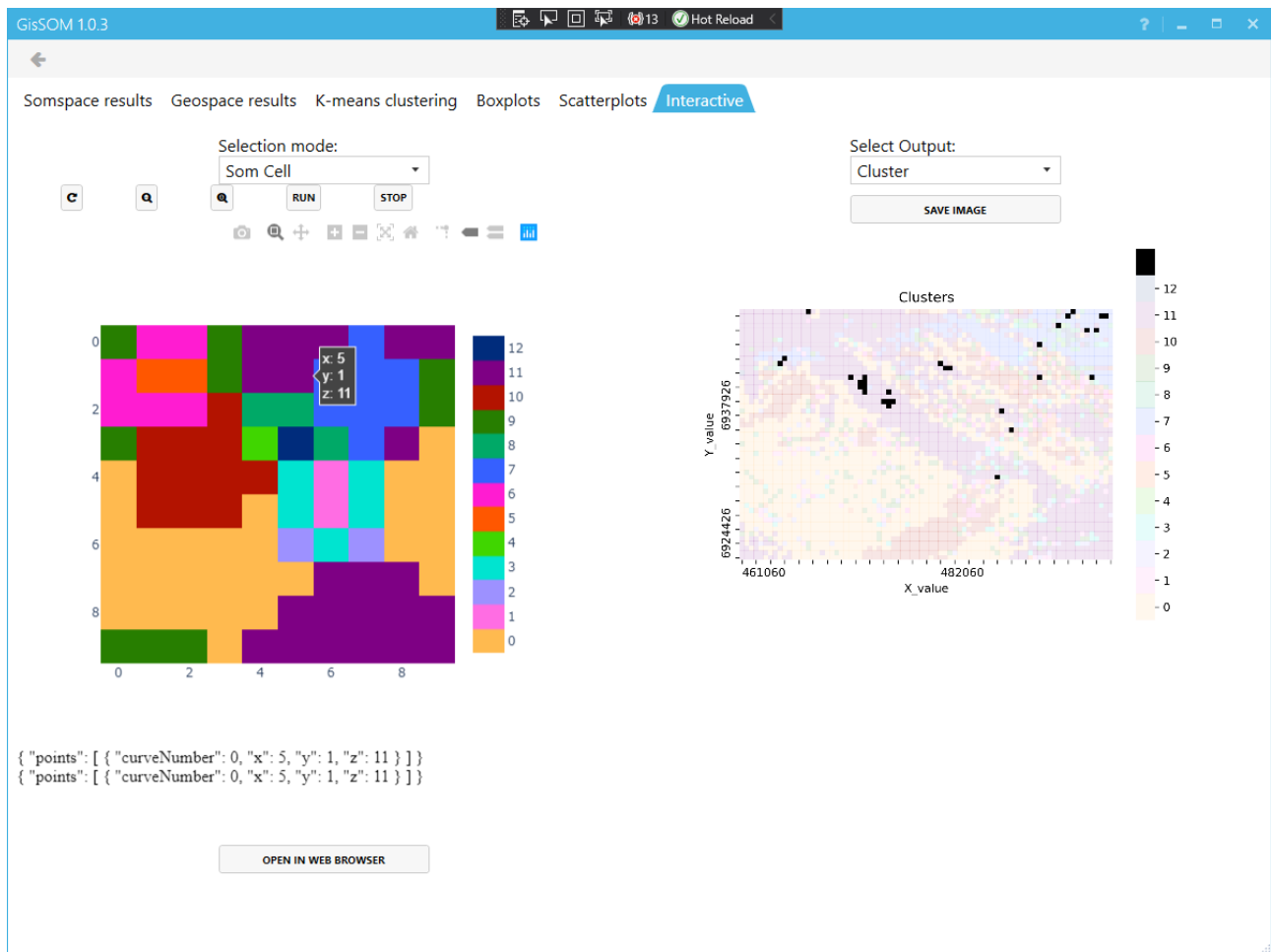
**Figure 14.** Scatterplots of selected variables

## Interactive plotting

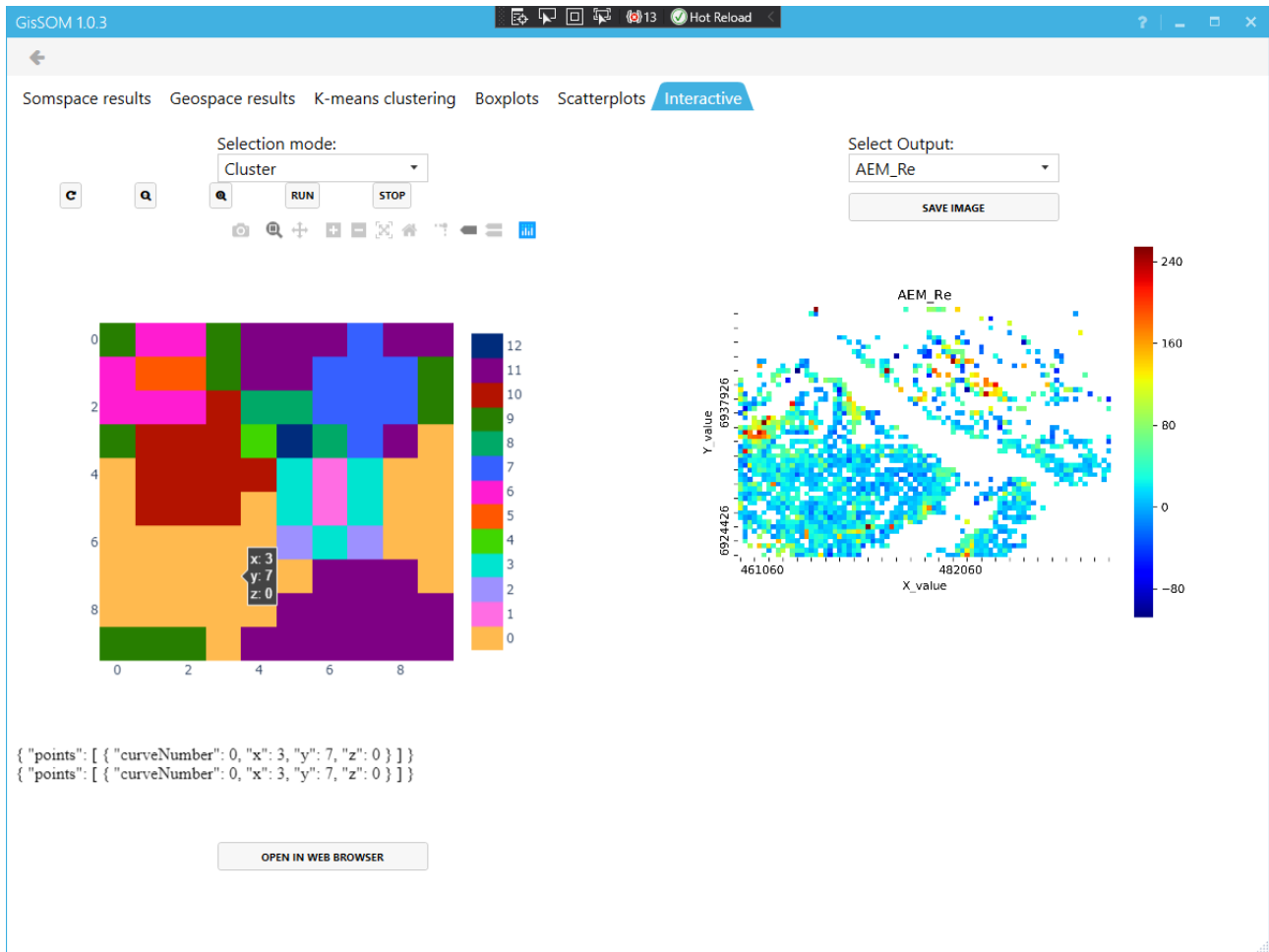
Hovering the mouse over the plot will display the x and y coordinates, and the number of the cluster (Figure 15). There are two selection modes: Cluster, which highlights an entire cluster on the geospacer map, and Som cell, which highlights one som cell. By left-clicking on any cell on SOM, the corresponding cluster or som cell will be highlighted on the geospace plot on the right side (). For the output, you can choose between clusters or any of the original data variables.



**Figure 15.** Interactive plot, selected cluster is drawn on the right-side image



**Figure 16.** Interactive plot highlighting a single SOM cell



**Figure 17.** Selected cluster highlighting a variable of the original data

In some cases, the built-in web browser that GisSOM relies on, might not display the interactive plot. In this case you can try to click the “Refresh this page” –button, or if this doesn’t work, you can open the plot in a web browser. The interactive plot should work on any up-to-date version of most modern web browsers (Edge, Chrome, Firefox...), and can be opened either by clicking the “Open in web browser”-button in the bottom of the interactive window, or by manually navigating to the web address “localhost:8050” (). The plot will work the same in both the GisSOM window and the web browser, where clicking on any SOM cell in the interactive plot will highlight the corresponding cluster on the main windows’ interactive geospace cluster plot.

## 4 REFERENCES

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224-227.

Deliverable 4.11 Appendix 1: Technical Specification – *nextsomcore*

Deliverable 4.11 Appendix 2: Technical Specification – *GisSOM*.

Deliverable 4.12: SOM tool for advangeo® (under preparation, due in M18)

Deliverable 4.13: SOM tool for ArcGIS (under preparation, due in M18)

Kohonen T., 2001. Self-organizing maps, Third Extended Edition, *Springer Series in Information Sciences*, 30.

Wittek, P, Gao, S. C., Lim, I. S., Zhao, L. (2017). Somoclu: An Efficient Parallel Library for Self-Organizing Maps. *Journal of Statistical Software*, 78(9), 1-21.