



GisSOM User Guide

Contents

Documentation page

1	introduction	1
1.1	Self-organizing maps and k-means	1
2	Installation	2
2.1	Installation requirements	2
3	using the software	2
3.1	Load the data	2
3.2	Data preprocessing	5
3.3	Choose SOM and k-means parameters	11
3.3.1	<i>Basic Parameters for SOM</i>	11
3.3.2	<i>Parameters for k-means</i>	13
3.3.3	<i>Output Folder</i>	14
3.3.4	<i>Advanced Parameters</i>	15
3.3.5	<i>Run and visualize results</i>	16
3.4	Results	16
3.4.1	Results in SOM space	18
3.4.2	Results in geospace	19
3.4.3	K-means clustering	20
3.4.4	Boxplots	22
3.4.5	Scatterplots	23
3.4.6	Interactive	24
3.5	Visualizing previous results	28
4	references	30

1 INTRODUCTION

The purpose of this document is to explain how to install, use and read the results of the *GisSOM* software developed initially in the European Union funded H2020 project NEXT (2018-2021), and upgraded in the EIT Raw Materials funded project DroneSOM (2022-2024). GisSOM is developed for performing self-organizing maps and k-means clustering for multivariate data. Before the SOM computation, the software provides means to do simple preprocessing to the input data and select variables for clustering. In addition to generation of SOMs from the input data, the SOM and k-means results for spatial datasets can be visualized in the original spatial coordinate system as well. GisSOM saves the mapping from the SOM coordinates to the original spatial coordinates for spatial datasets as csv tables. 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 a separate technical specification document.

1.1 Self-organizing maps and k-means

Self-organizing map (SOM) is an unsupervised artificial neural network that arranges a set of n-dimensional vectors to a usually two-dimensional grid of neurons (Kohonen, 2001). The usability of SOM comes from its topology preserving nature, i.e., similar data vectors are assigned to SOM neurons, sometimes called nodes or cells, that are close to each other.

SOM computation is carried out using the somoclu package (Wittekk 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 in feature space 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 defines the amount which the weights are updated and often decreases over time. The neighborhood function ensures that the update is smaller the farther away the neuron is from the BMU in SOM space. Also the neighborhood size usually decreases as a function of time.

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 neuron. In GisSOM, computation of only quantization error is implemented so far. Quantization error represents the deviation of each data vector from its corresponding BMU vector.

Although SOM can be considered as a clustering method itself, the number of neurons in a SOM is generally too large for practical data classification. To produce a smaller number of clusters, GisSOM applies k-means clustering to the SOM neurons. K-means is a very basic clustering method that tries to

December 5, 2022

minimize variances within clusters without considering the relation between or similarity of different clusters. The algorithm is iterative; it assigns observations to the closest cluster centroid and recalculates the centroids. This is repeated until no updating happens.

In GisSOM, the user provides the minimum and maximum number of clusters for which k-means clustering is computed. Because 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. 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 the executable SomUI.exe, which is in the root of the installation directory .../GisSOM. This will open a wizard-style window (Figure 1) 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) on 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 variables, but in the case of GeoTIFF files, multiple files can be given, each containing a single data variable. When opening a CSV file, make sure the file is not open in Excel or any other spreadsheet-editing application, because then it can't be accessed.

Input data options are

- **CSV Grid:** Use this for a table representing gridded spatial data. This works only if there are no missing full rows or columns in the grid. Individual points can be missing anywhere on the grid. The format is 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 must have a non-empty header. The headers can contain numbers, but the first character of a header can't be a number. Also, spaces (), percentage signs (%) and double quotation marks ("") are removed from headers. If whole columns and rows are missing (NoData), CSV Scatter format is needed. See also the example testdata file in .../GisSOM/TestData.

December 5, 2022

- **CSV Scatter:** Use this for spatial data that is not evenly spaced in a grid and for non-spatial data. The format requirements of CSV Grid apply to CSV Scatter as well. If there are only individual missing values or missing values in the sides and center in a gridded spatial dataset, CSV Grid input type can be used (Figure 2a), but if whole columns and rows are missing, CSV Scatter is needed (Figure 2b). CSV Scatter is useful, for instance, for geochemical spatial data, where the samples are collected at irregular spatial intervals (Figure 2c). The main drawback of the CSV Scatter input is that it makes it slower to produce the geospace result images than the CSV Grid input.
- **TIF:** A georeferenced single-band raster data format. In case of multi-band GeoTIFF files, each band within the raster should either be saved as individual single-band raster file or all the band values should be sampled and saved in CSV Grid format. Note that when using raster data, the alignment and size of the pixels and the NoData-value need 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. If entire rows or columns are NoData, CSV Scatter input format must be used. If whole columns and rows are missing (NoData), CSV scatter format is needed.
- **Old results:** Use this option if you want to visualize existing SOM results from GisSOM software. Navigate to the output folder GisSOM and select the result subfolder (Out_*). Make sure you haven't changed the file names or paths of the results. See section 3.5 for more information.

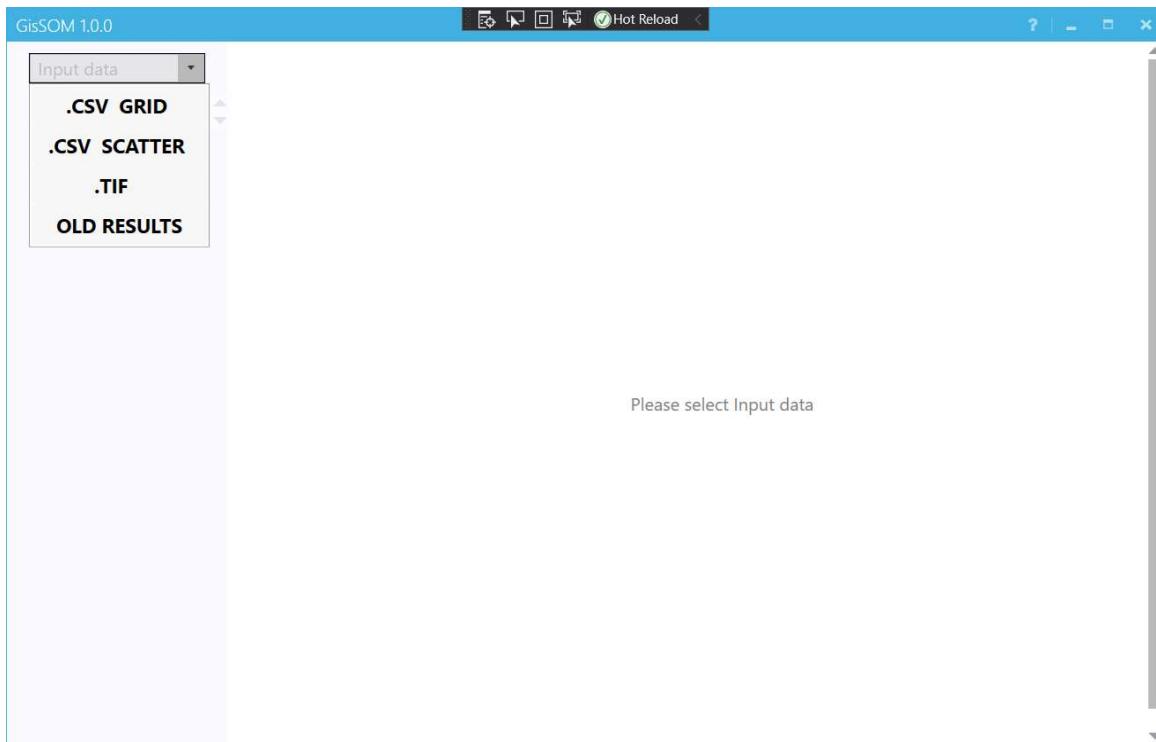


Figure 1. Selecting the data format. The data processing options will be visible only after the input data is loaded (see Figure 3).

December 5, 2022

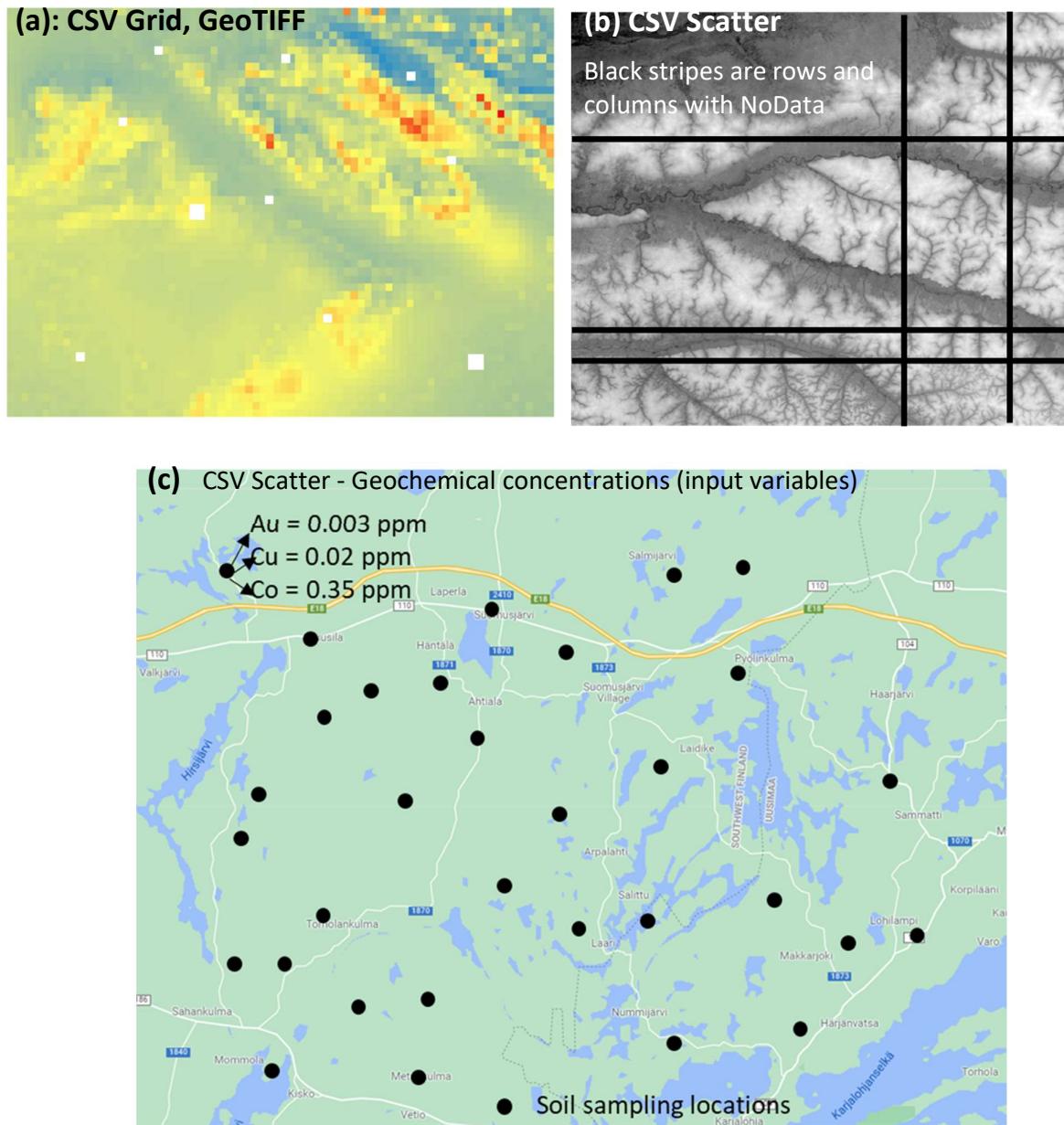
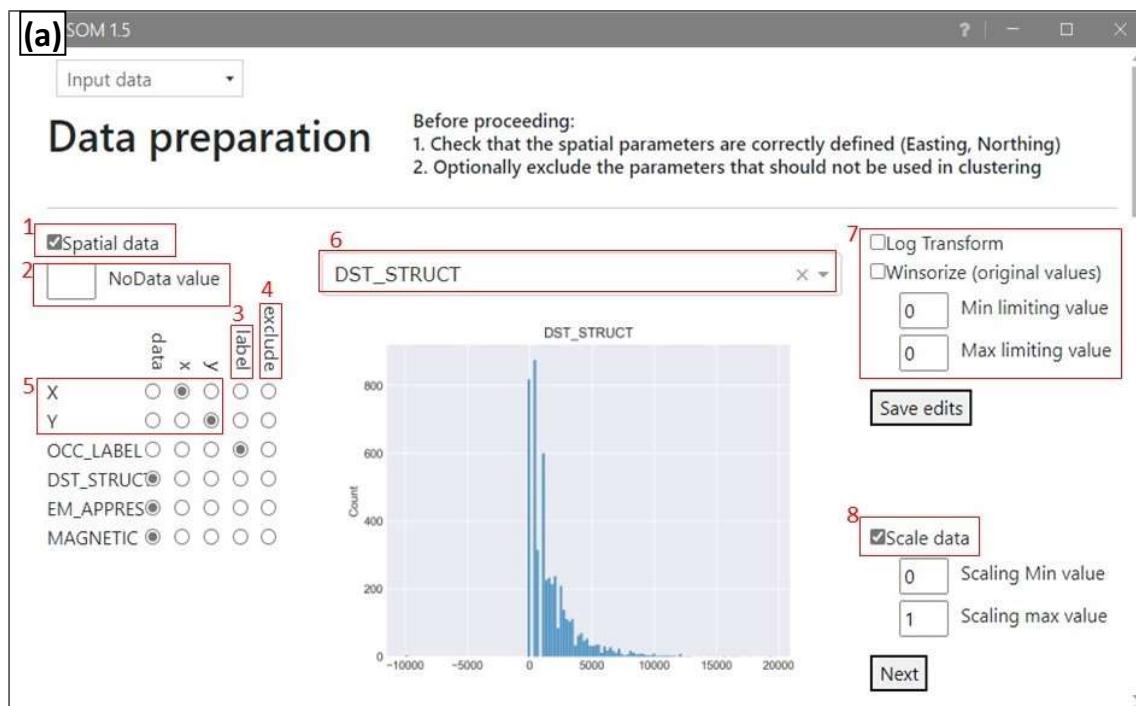


Figure 2. Example of grid type and scattered data points. The data in (a) can be used as grid data, as there are no missing columns or rows; except a few pixels (white pixels) where it is NoData. The data in (b) has several rows and columns missing (black stripes), e.g.: satellite imagery, so CSV scatter needs to be used, otherwise the result plots will misrepresent the data. CSV scatter can also be used in case of data collected from scattered locations as shown in (c) typical for, e.g., geochemical data.

December 5, 2022

3.2 Data preprocessing

Once you have selected the input data file(s), the GisSOM window shows you all the data variables (Figure 3). In the case of CSV input, also the possible spatial variables and labels are shown. In the next step, you need to arrange the data before it can be used in SOM. You can exclude any variable you don't want to use in SOM, and possibly transform data. In the case of CSV spatial data, you need to select the correct North and East coordinates and define the possible label variable; when defined, these are automatically excluded from SOM computation. 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.



December 5, 2022

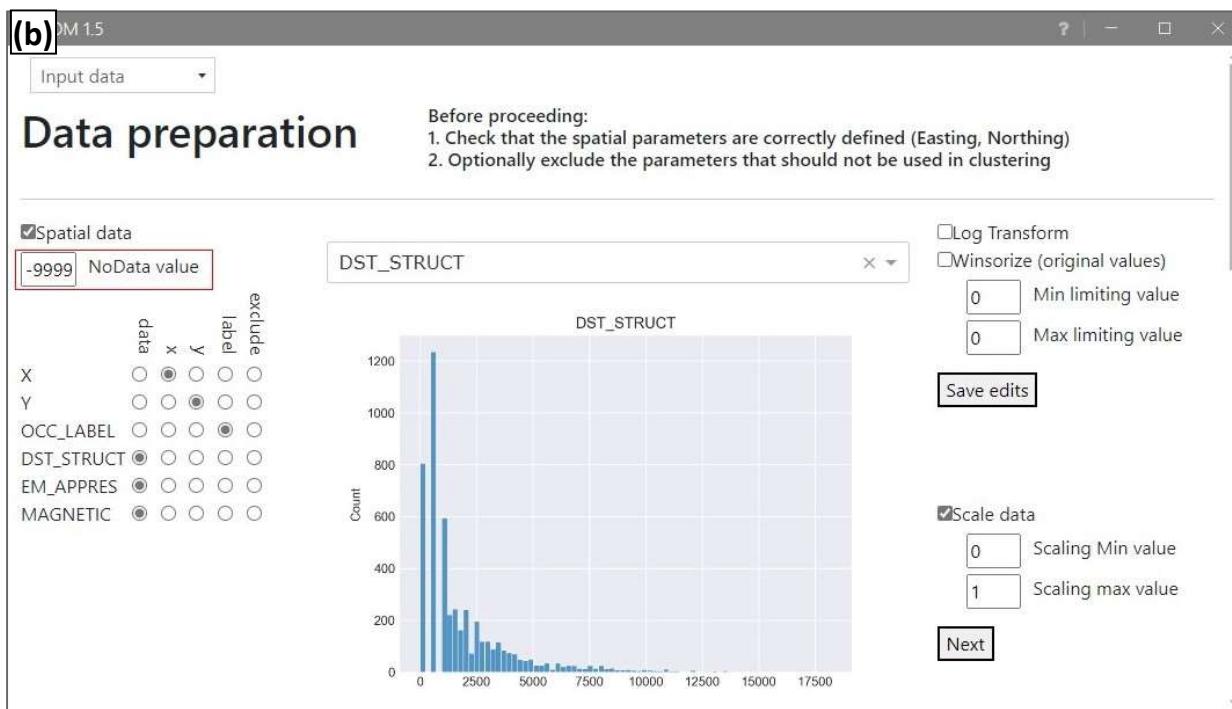


Figure 3. a) GisSOM Data preprocessing interface after data is loaded. Steps after loading data: 1. For spatial data, check the *Spatial data* checkbox, 2. Specify NoData value for CSV input (see figure b), 3. Specify the label column if such exists, 4. Optionally, specify columns to be excluded, 5. Mark x and y columns for spatial data, 6. Draw histogram of the selected variable, 7. Data preprocessing options and 8. Scale all variables to the same interval. b) Histogram is updated based on the given NoData value.

The following steps are required (Figure 3):

- For spatial data, select the North and East coordinates by checking the “x” and “y” radio buttons for the correct variable from the left. This should not be required for georeferenced raster data (GeoTIFF), but if the spatial variables are not marked as x and y, it is safest to manually assign them. If you are not using spatial data, untick the checkbox “*Spatial data*” on the right side. If you have labels in your data, i.e. a variable that represents true values, mark them using “*Label*” radio button.
- Exclude the data variables that you don’t want to be used in SOM by choosing “*Exclude*” for the correct variables on the left side. Note that geographic coordinates are automatically excluded if you define them as “x” and “y”, as you normally do not want to use them in the SOM as variables. Also, the variable marked as “*label*” is automatically excluded.
- If dataset contains a numeric value signifying NoData 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 (Figure 3). GisSOM automatically screens for non-numeric NoData values, so those do not have to be provided.
- The histogram for the selected variable is displayed in the center. This helps to analyze the distribution of the selected variable’s data values. Study the distribution of the variables that

December 5, 2022

you want to use in SOM computation using their respective histograms. If the displayed histogram has a long tail in the high end, you can apply logarithmic transformation ("Log transform") or if both tails are long, limit extreme values ("Winsorize", Figure 4) to emphasize differences near the peak of the distribution. Click the "Save edits" button to apply the change and see it in the histogram.

- Logarithmic transform is carried out by first shifting all the variable values to the positive range. Then, a natural logarithm is applied to the values and SOM computation is performed using the transformed values. After the SOM computation, 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 variable 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 original data values, irrespective of the values shown in the histogram, i.e., even if you apply logarithmic transformation first, the original data values (and not the log-transformed data values) will be winsorized.

Either of the two preprocessing options can be applied to one variable, not both. If in case the user wishes to undo the logarithmic transformation or winsorization, then uncheck the corresponding option and then click "Save edits" to refresh and visualize the original data distribution.

- Often it is important to scale the data variables into a comparable range of values for SOM computation. Exception is, for instance, spectral reflectance data, where all the variables already represent the same range. If the "Scale data" checkbox is checked, each variable is scaled by default to the range of [0,1], after pressing the "Next" button. If the user wants to weigh data variables differently, they can provide variable specific ranges. To define the range for a specific variable, select the variable from the dropdown above the histogram, define the minimum and maximum values for scaling and press the "Save edits" button (Figure 5). All the variables for which a range was not defined are scaled to range [0,1]. 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. Histogram on the data preparation interface shows unscaled data values.

- Click "Next" to proceed to the next step

December 5, 2022

(a) GisSOM 1.5

Data preparation

Before proceeding:

1. Check that the spatial parameters are correctly defined (Easting, Northing)
2. Optionally exclude the parameters that should not be used in clustering

Spatial data
-9999 NoData value

	data	x	y	label	exclude
X	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Y	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
OCC_LABEL	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
DST_STRUCT	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
EM_APPRES	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
MAGNETIC	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

MAGNETIC

Count

0 -1000 -500 0 500 1000 1500 2000 2500

Log Transform
 Winsorize (original values)
0 Min limiting value
0 Max limiting value
Save edits

Scale data
0 Scaling Min value
1 Scaling max value
Next

(b) GisSOM 1.5

Data preparation

Before proceeding:

1. Check that the spatial parameters are correctly defined (Easting, Northing)
2. Optionally exclude the parameters that should not be used in clustering

Spatial data
-9999 NoData value

	data	x	y	label	exclude
X	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Y	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
OCC_LABEL	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
DST_STRUCT	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
EM_APPRES	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
MAGNETIC	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

MAGNETIC

Count

0 5.5 6.0 6.5 7.0 7.5 8.0 8.5

Log Transform
 Winsorize (original values)
0 Min limiting value
0 Max limiting value
Save edits

Scale data
0 Scaling Min value
1 Scaling max value
Next

December 5, 2022

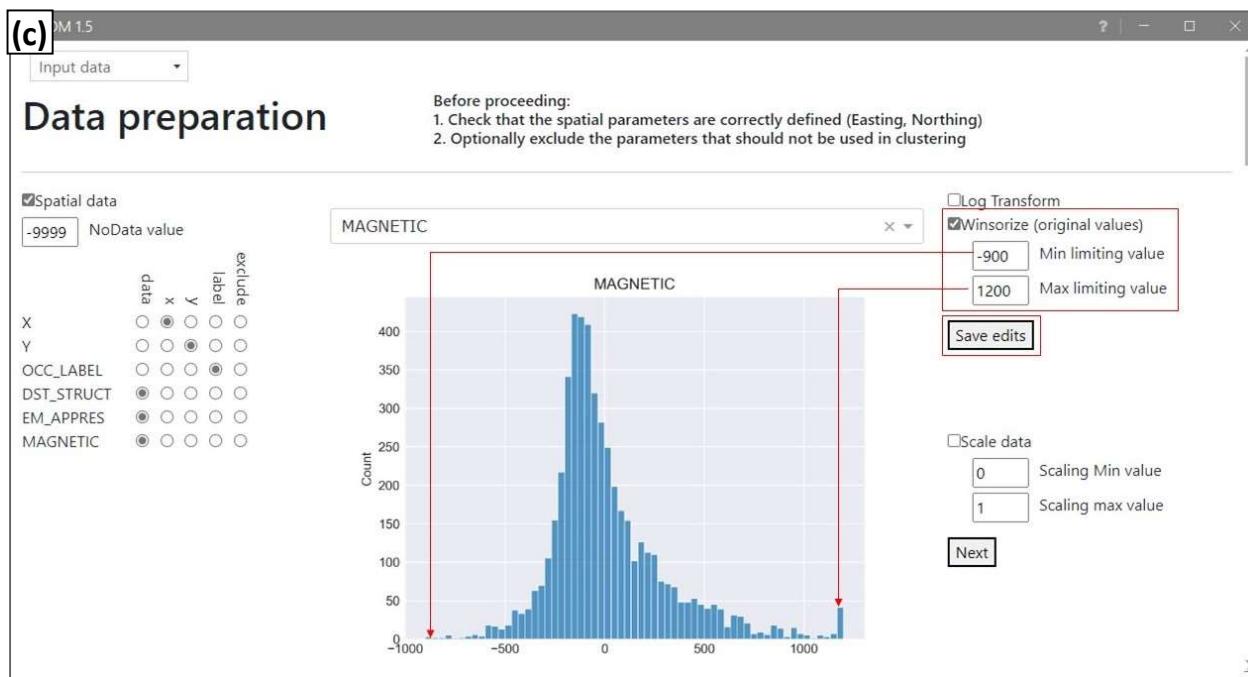


Figure 4. Data processing and visualization. a) No preprocessing, the histogram displays original data values, b) Log transformation of original data values, and c) Winsorizing of the original data values.

December 5, 2022

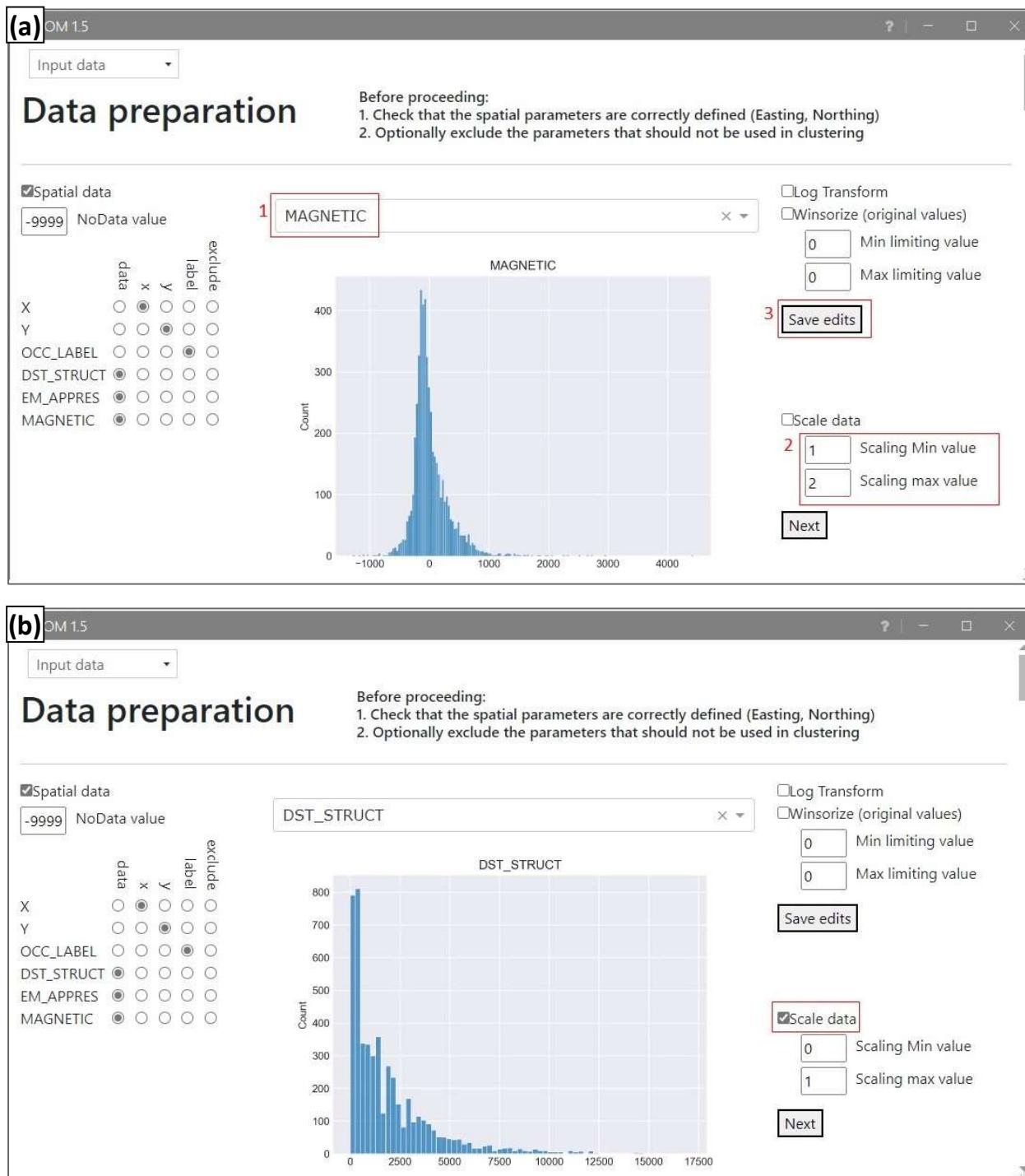


Figure 5. Defining parameters for scaling variables. a) If the default range [0,1] is not applicable for some variables: 1. Select the variable for which the range is to be defined, 2. provide minimum and maximum values and 3) press the “Save edits” button to save the

December 5, 2022

range. b) Before proceeding from the Data preparation page, check the “Scale data” checkbox. All the variables for which the range was not specified, are scaled to range [0,1].

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 6). 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.

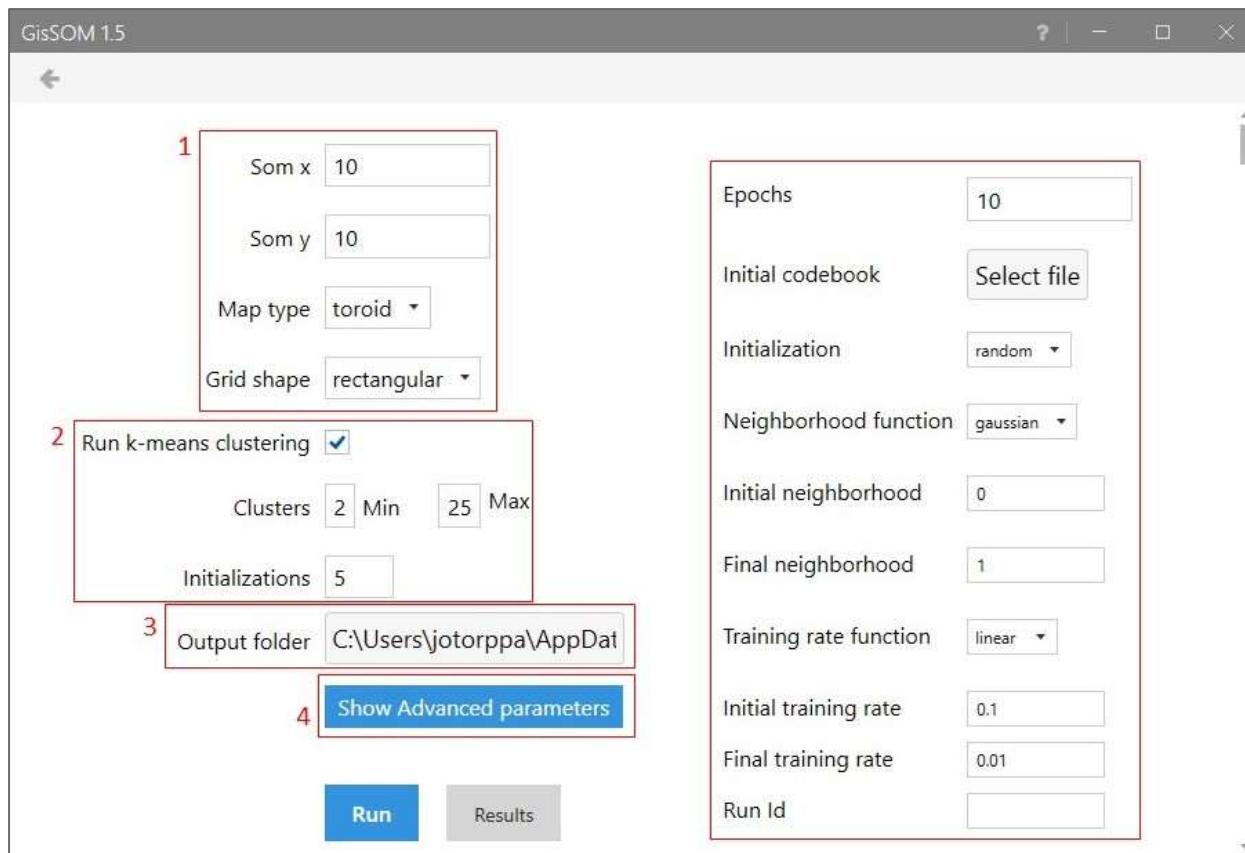


Figure 6. Defining input parameters for SOM and k-means. 1. Basic input parameters for SOM computation, 2. Input parameters for k-means computation, 3. Output path and 4. Button for showing the advanced SOM computation parameters.

3.3.1 Basic Parameters for SOM

- Choose the size of the SOM using “Som x” and “Som y” (Figure 6). 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. A large SOM reveals more details in the input data with the cost of computing time.

December 5, 2022

- Map type: topology of the map, accepted values are “toroid” or “sheet” (Figure 6). 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 7.

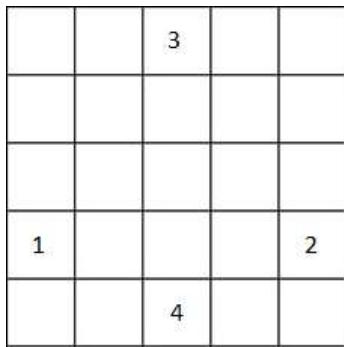


Figure 7. Neighbours in sheet and toroid type maps. For toroid, 1 and 2 as well as 3 and 4 are neighbours; for sheet they are not.

Using sheet type often causes large and small values to be clustered in corners whereas toroid doesn't have this issue. Using toroid type often causes a single cluster to be split in multiple parts in visualization, as the clusters continue from one side to the opposite one. These cases are visualized in Figure 8.

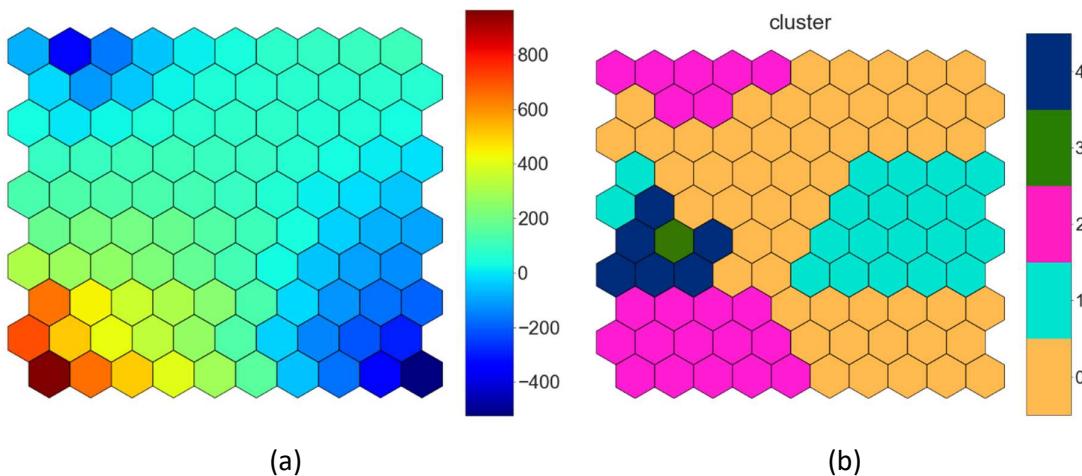


Figure 8. Sheet type causing extreme values to be piled in the corner (a) and toroid type causing a cluster to be split on opposite sides (b).

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

December 5, 2022

3.3.2 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 (Figure 6).
- If you run k-means, you need to select the minimum and maximum number of clusters (Figure 6). 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 (Figure 6). The default is 5. K-means is sensitive to the initialization, so multiple different initializations should be used. The software runs k-means using these different initializations and chooses the most optimal based on the smallest *Davies-Bouldin index*.

December 5, 2022

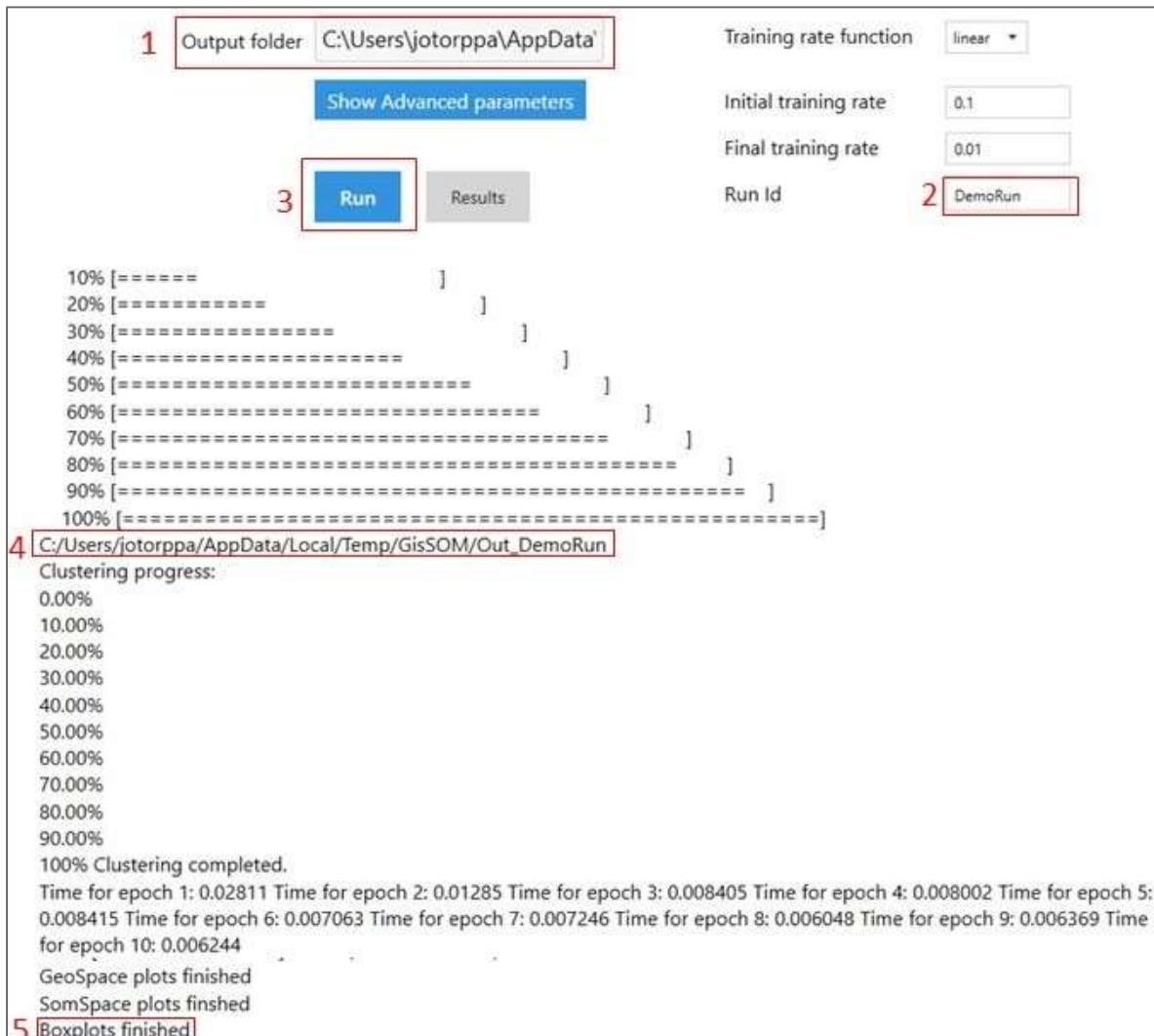


Figure 9. Computation progress and output folder. 1. Define the output path, 2. Define the name of the subfolder where output from the active SOM run is written and 3. “Run” button for starting the computation. 4. Output path is shown in the printed output. 5. Computations are finished, when boxplots have been drawn.

3.3.3 Output Folder

- The default output folder location where all the results will be saved is “C:\Users\<YourUserName>\AppData\Local\Temp\GisSom\”.
- The default output folder can be changed using the *Output folder* textbox (Figure 9).
- A new folder “GisSOM” is created in the path defined by the user. The results for each instance are saved within this “GisSOM” directory, in a subdirectory, the name of which starts by “Out_” followed by the *Run Id* (Figure 9). *Run Id* is by default the date and time and SOM size-shape

December 5, 2022

parameters unless the it is provided in the Advanced Parameters (see section 3.3.4). The timestamp is in the format “YYYYMMDDHHMMSS” (year-month-day-hour-minute-second).

- The final structure is *Output folder\GisSOM\Out_Run Id* where the text in italics can be defined by the user.

3.3.4 Advanced Parameters

Advanced parameters are shown in Figure 6.

- *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 number of epochs.
- *Initial codebook*: 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 result subfolder, under the main “GisSOM” output directory. 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. With random initialization, the initial codebook will be filled with random numbers ranging from 0 to 1. “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 neighborhood* 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 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”.

December 5, 2022

- *Run Id:* Use *Run Id* to change and define the name of the output directory within the “GisSOM” folder. The results will be saved in the new directory with the name specified in *Run Id* prefixed by “Out_”. (Figure 9)

3.3.5 Run and visualize results

- Click “Run” to run the software
- The log is generated and after successful SOM run, “Som run complete” notification is displayed.
- Log also shows the progress of k-means clustering and drawing of images.
- Click “Results” to visualize and analyze the results.

3.4 Results

In the last step, you can see the results. In the user interface there are six different views for different types of results: “Somspace results”, “Geospace results”, “K-means clustering”, “Boxplots”, “Scatterplots” and “Interactive” (Figure 10).

The results are also written in the filesystem. The main output folder (defined by the default timestamp or the *Run_Id* parameter (Sec 3.3.4)) includes the SOM space and geospace results (“result_som.txt/.csv” and “result_geo.txt/.csv”) and information on the run (“RunStats.txt”) as text files, and dictionary files (“som.dictionary”, “cluster.dictionary”). The “result_som.*” contains the SOM codebook vectors and other SOM neuron related information. The SOM codebook vectors represent the values of each SOM cell in the input data variables. The “results_geo.*” contains SOM results mapped to the spatial reference frame of the input data.

Subfolders in the main results folder, corresponding to each tab, can be accessed by pressing the “Show results in filesystem” button on the respective tab (except K-means clustering, from which results are not output as files). Folders contain figures and/or data for geospace and SOM space results, boxplots, scatterplots, data preparation, original data, and the interactive plot. If input data is given in geoTIFF format, results in geospace are also given in geoTIFF format in addition to the results_geo.* files. The folder structure is visualized in Figure 11.

December 5, 2022

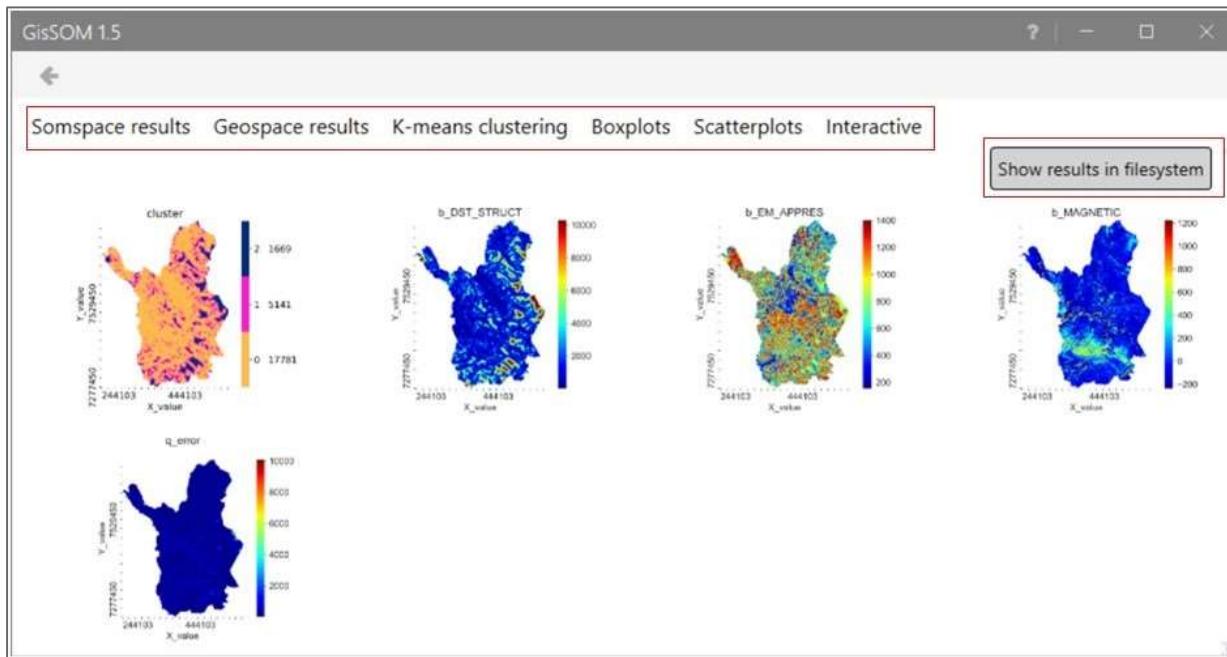


Figure 10. Tabs for viewing different types of results (Somspace results, Geospace results, K-means clustering, Boxplots, Scatterplots and Interactive) and the button “Show results in filesystem” for showing the folder where results have been written.

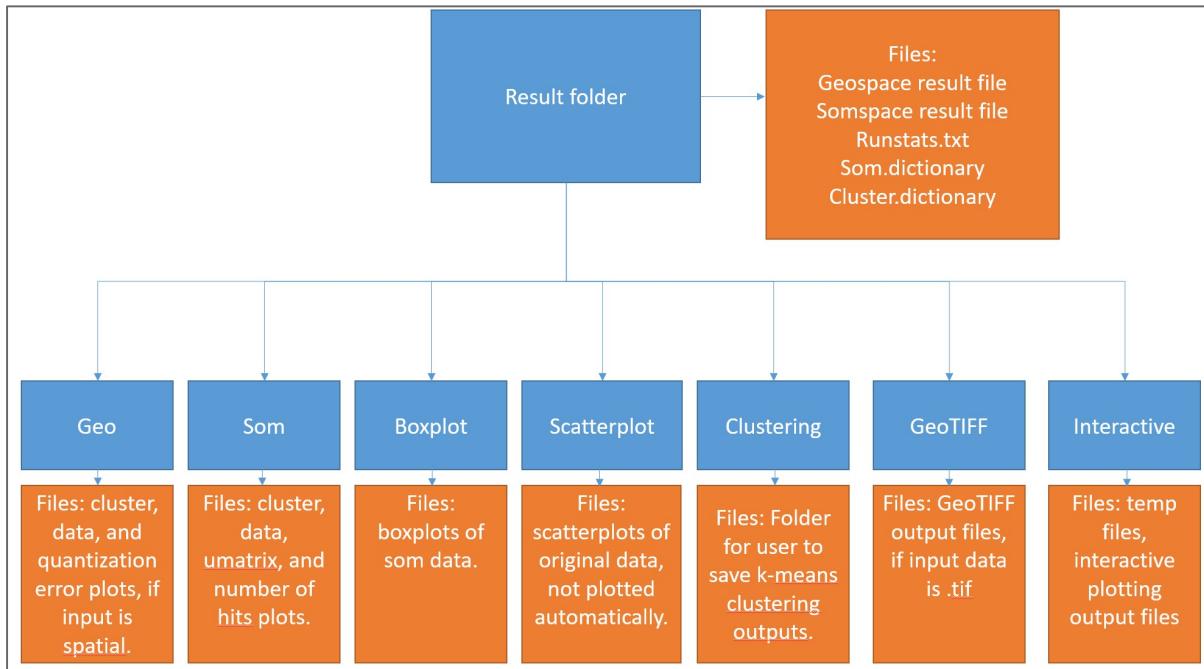


Figure 11. Structure of the result folders. Blue boxes represent folders and orange boxes represent files. For labelled input data in the csv format, a ‘labels.csv’ file is also generated in the main Result folder.

December 5, 2022

3.4.1 Results in SOM space

These images (Figure 12) show the resulting SOM, color coded using various parameters:

1. K-means cluster (Figure 12a). The left-most numbers on the legend represent the number of the cluster, and the numbers on the right represent number of data points per cluster (Figure 12b).
2. If labels are provided in the input data, the k-means cluster image shows the labels on the SOM cell. The labelled SOM nodes are indexed so that each combination of labels assigned to a SOM neuron gets a unique index. For instance, SOM cells with label index ‘1’ in the k-means cluster image contain the one data point with label ‘1’. SOM cells with label index ‘2’ contain two labelled data points each with label ‘1’.
3. Value of each codebook vector element (corresponding to data variables), Figure 12d-f
4. U-matrix (the magnitude of the difference of the codebook vectors in neighboring SOM nodes), Figure 12g
5. Number of data points clustered in each SOM nodes, Figure 12h.

The “Add label data” (Figure 12i) 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”).

All results can be viewed in the containing folder using the “Show results in filesystem” option (Figure 12j)

December 5, 2022

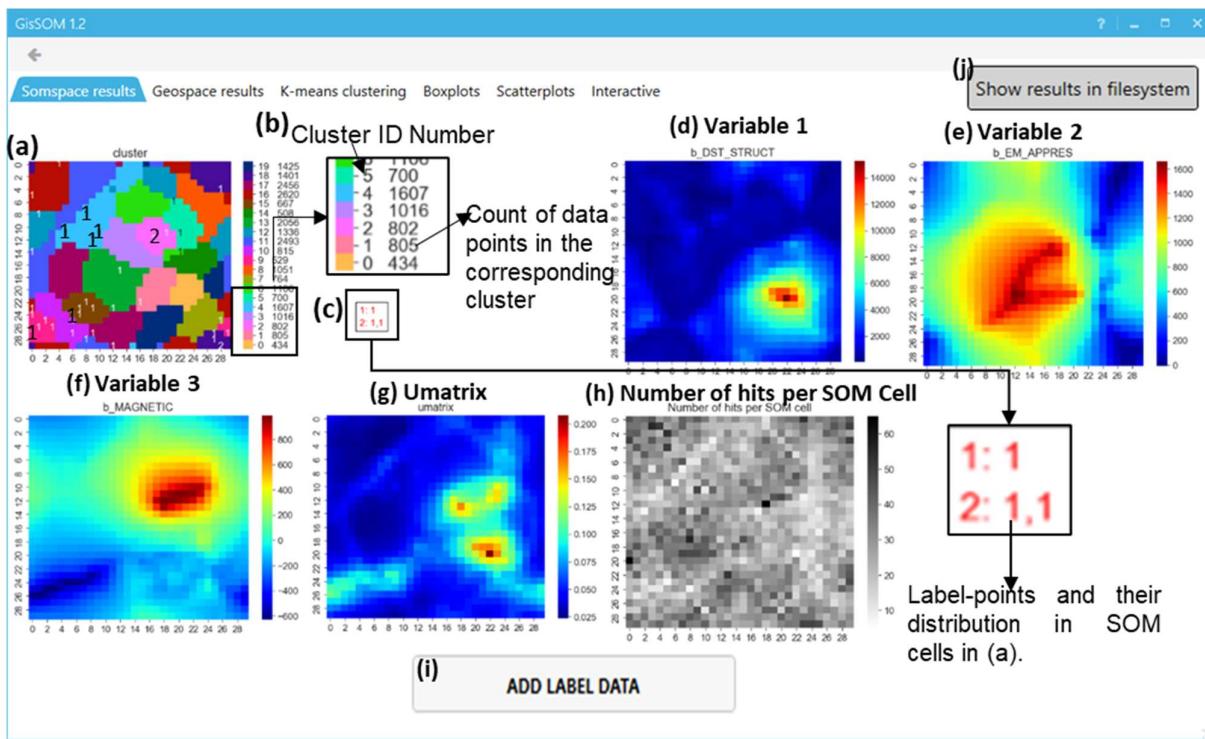


Figure 12. SOM space results. The frames (a)-(j) are described in Section 3.4.1.

3.4.2 Results in geospace

These images show the k-means clusters, the codebook vector elements (input variables) of the best-matching units and quantization error in geographical space (Figure 13a; b-d; and e, respectively). Quantization error is the difference between the original data variable values and the SOM codebook vector elements of the SOM node where the data point has been projected to. High quantization error values show outliers in the data.

For non-spatial input data, the “Geospace results” window remains empty. In such a case, the quantization error cannot be visualized, and it is reported only in the “RunStats.txt” -text file.

December 5, 2022

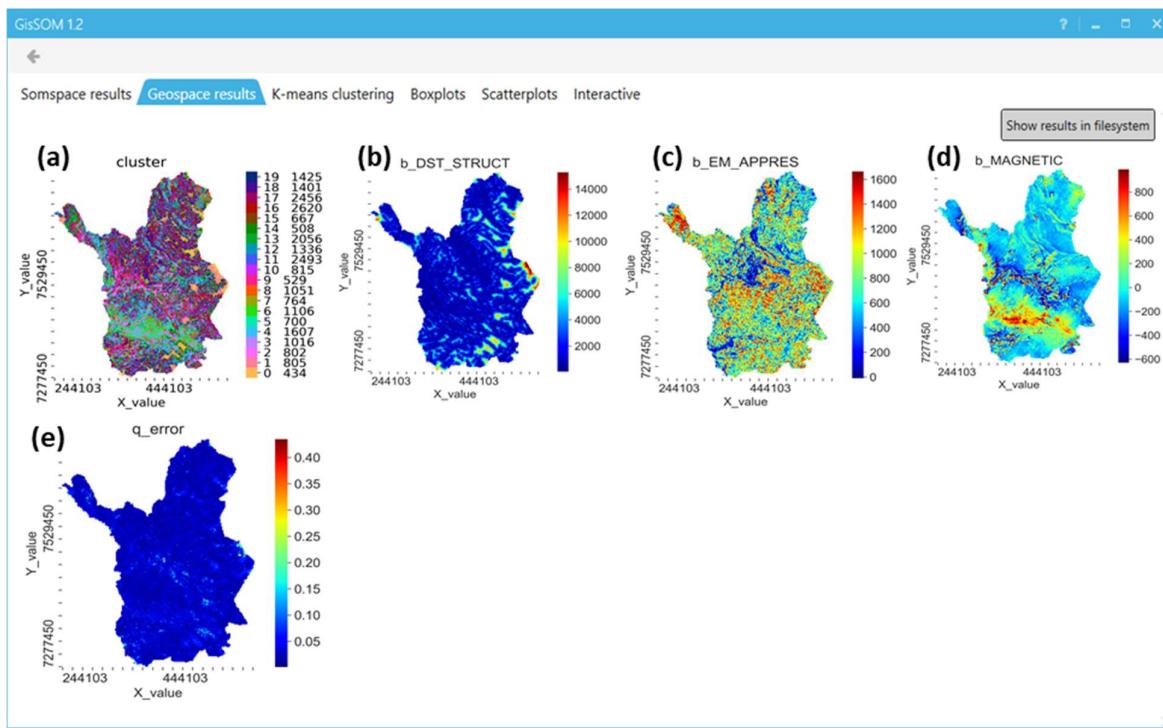


Figure 13. Geospace results.

3.4.3 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 14). You can run the clustering again with different minimum and maximum number of clusters or initializations (Figure 14a). It is advised to save the current clustering solution as a cluster dictionary file using the “Save Clustering Dictionary” option (Figure 14b). This dictionary file can be reloaded using the “Load Clustering Dictionary” option (Figure 14c). The plot in Figure 14d 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 (Figure 14f). This updates the results in the other tabs and the images will be redrawn with the selected number of clusters. These images will overwrite the previous images in the output folder.

December 5, 2022

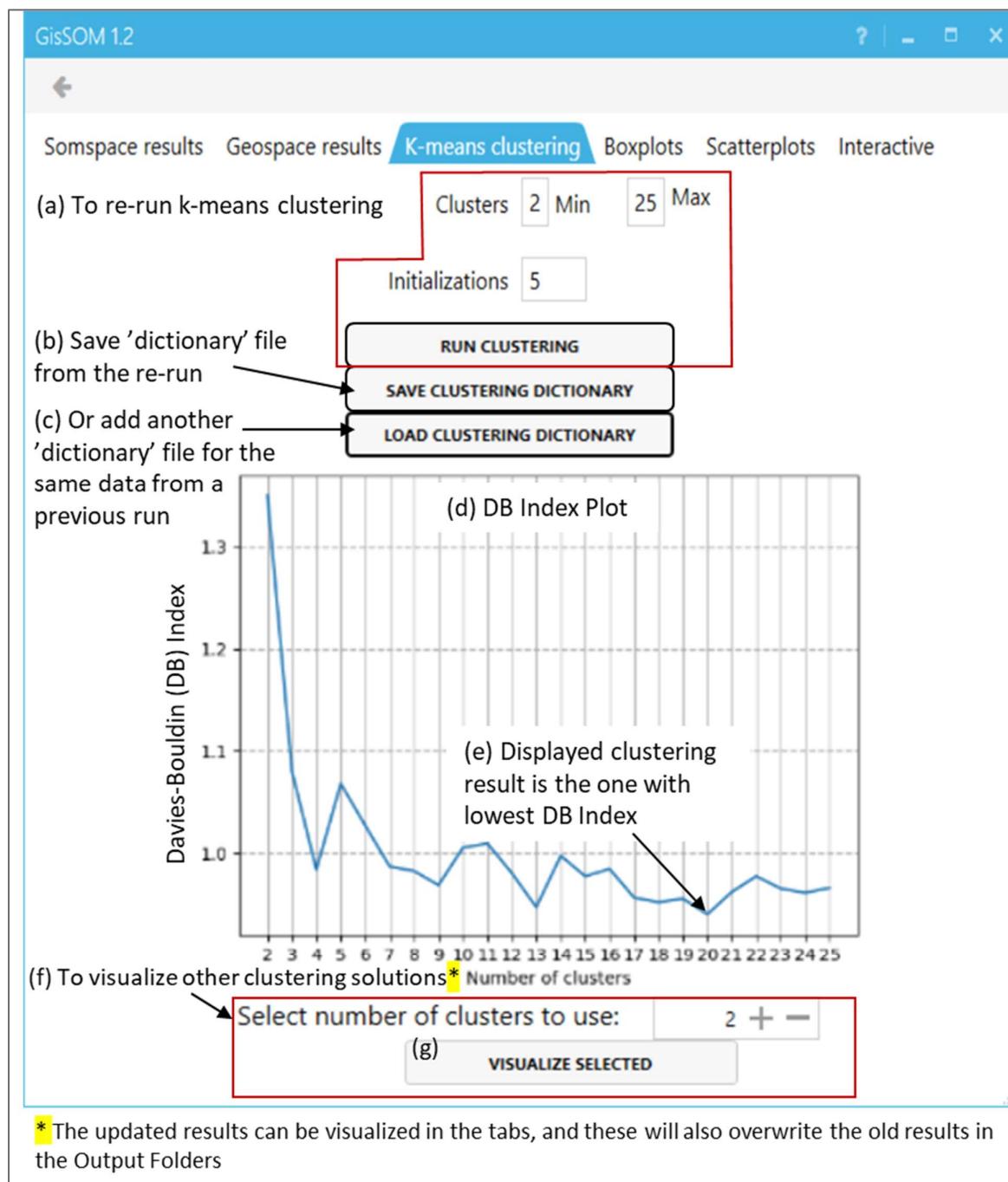


Figure 14. Clustering results.

December 5, 2022

3.4.4 Boxplots

These images show the boxplots for the SOM codebook vectors for each data variable. If the SOM result represents the data, the boxplots describe the distribution of original data values as well. The values are grouped by the k-means clusters (Figure 15). Color coding is the same as in the geospace and SOM space images.

The window includes one image for each data variable. In an image, there is one boxplot for each cluster. 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.

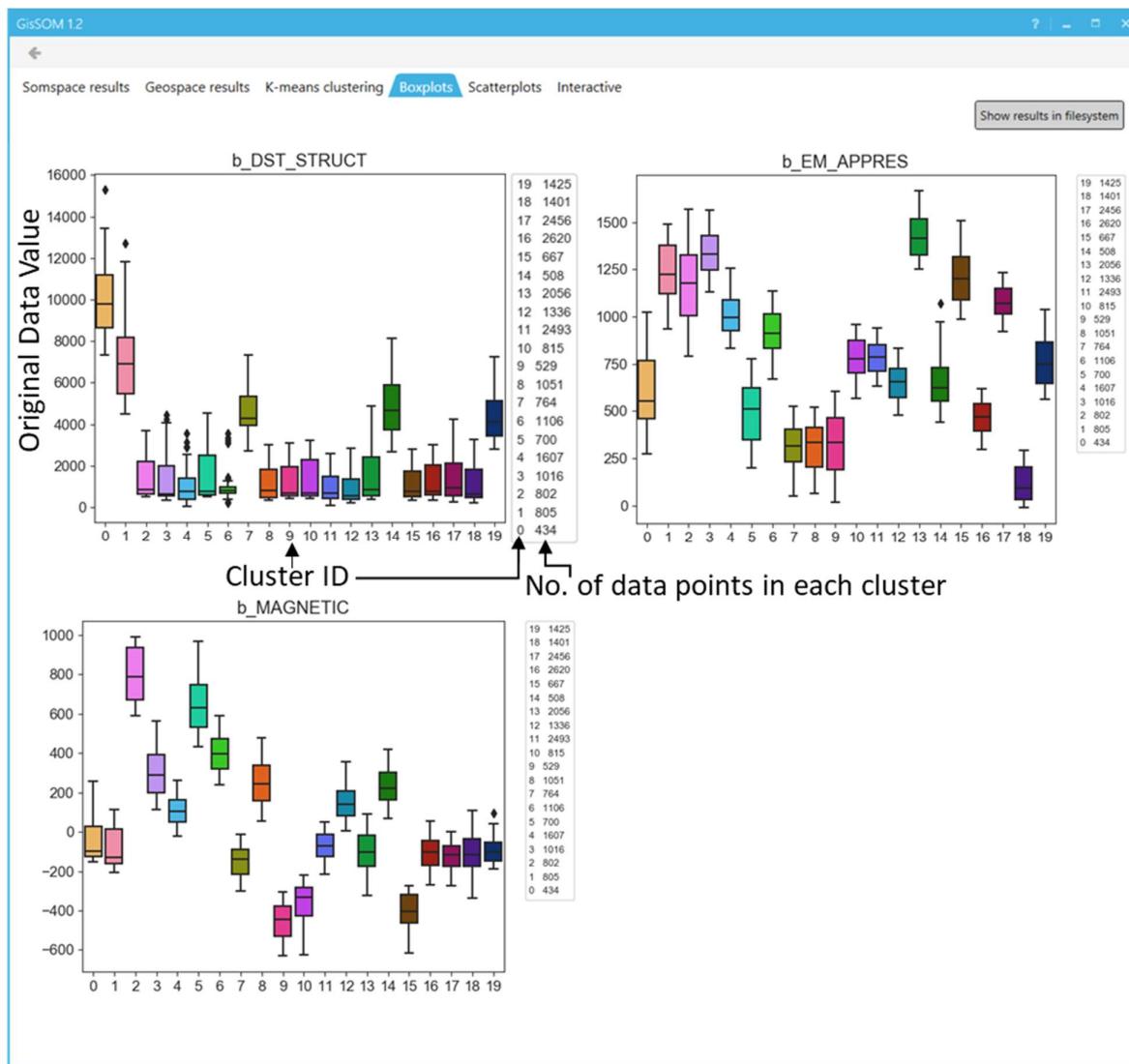


Figure 15. Boxplot results.

December 5, 2022

3.4.5 Scatterplots

Scatterplots are plotted for SOM codebook vectors. The initial Scatterplots tab is without any images (Figure 16). Scatterplots are not plotted or saved by default in the output folder. Because 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. This also saves the drawn scatterplots in the “Scatterplot” folder. Each image shows one pair of variables as a scatterplot (Figure 17) and each point in the scatterplot represents one SOM codebook vector. The dots are colored based on the cluster that the data point belongs to.

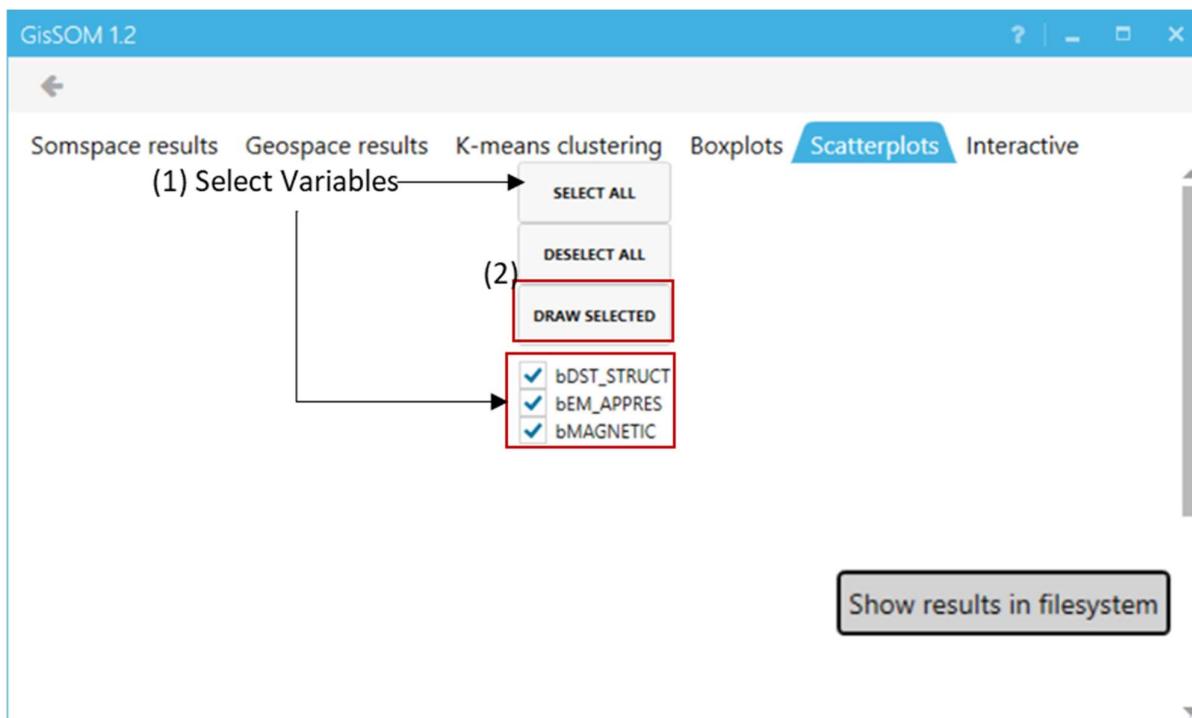


Figure 16. Scatterplot page after SOM calculation. No scatterplots have been drawn by default.

December 5, 2022

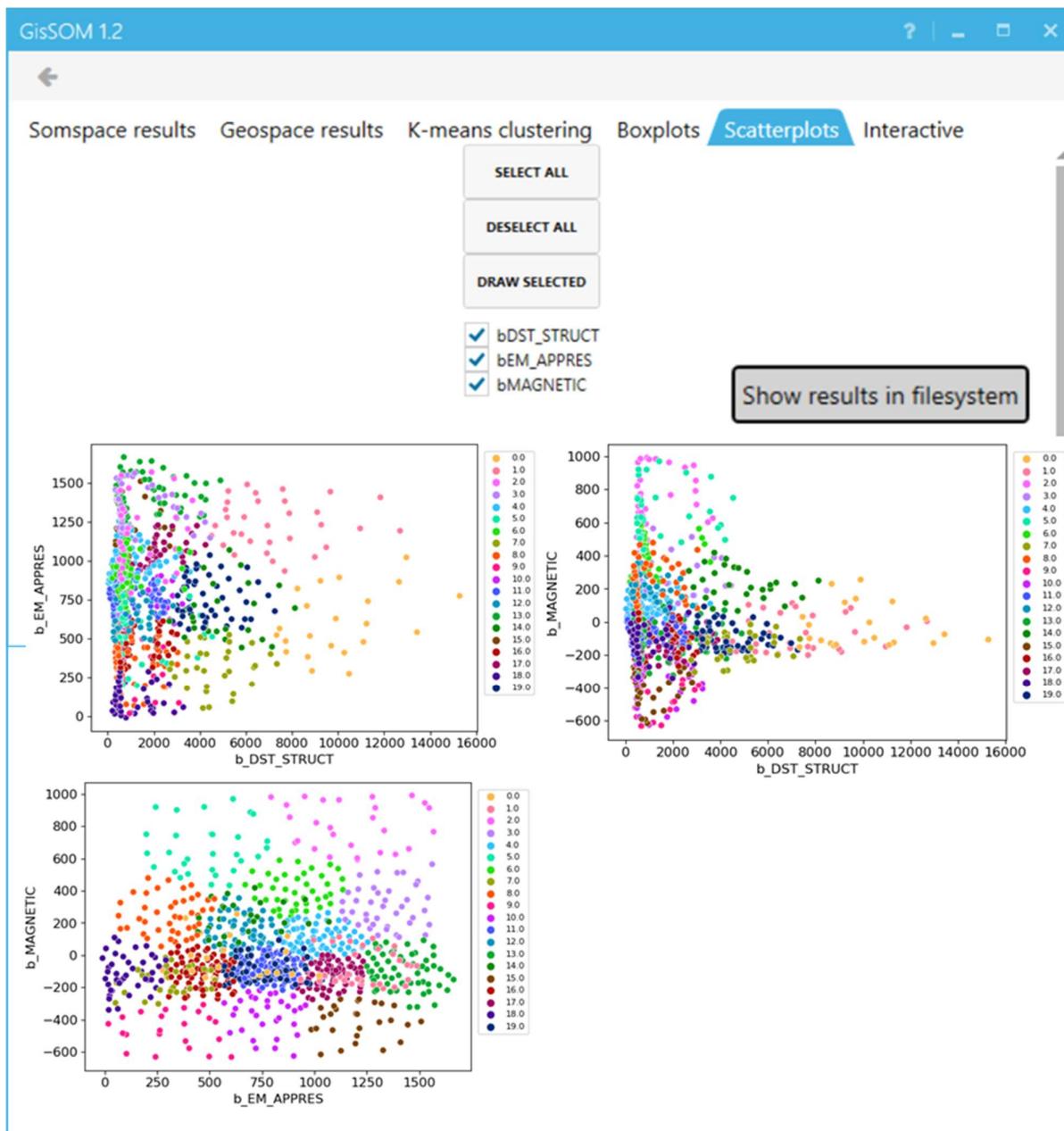


Figure 17. Scatterplots of selected variables.

3.4.6 Interactive

Hovering the mouse over the SOM grid plot will display the SOM x and y coordinates and the index of the cluster. There are two selection modes: “Cluster”, which highlights data points assigned to an entire cluster on the geospace map (Figure 18), and “Som cell”, which highlights data points assigned to one SOM neuron (Figure 19). By left-clicking on any cell on SOM, the data points assigned to the

December 5, 2022

corresponding cluster or SOM node will be highlighted on the geospace plot on the right side. As the geospacer image, you can choose between clusters or any of the original data variables (Figure 20).



Figure 18. Interactive plot, selected cluster is drawn on the right-side image.

December 5, 2022

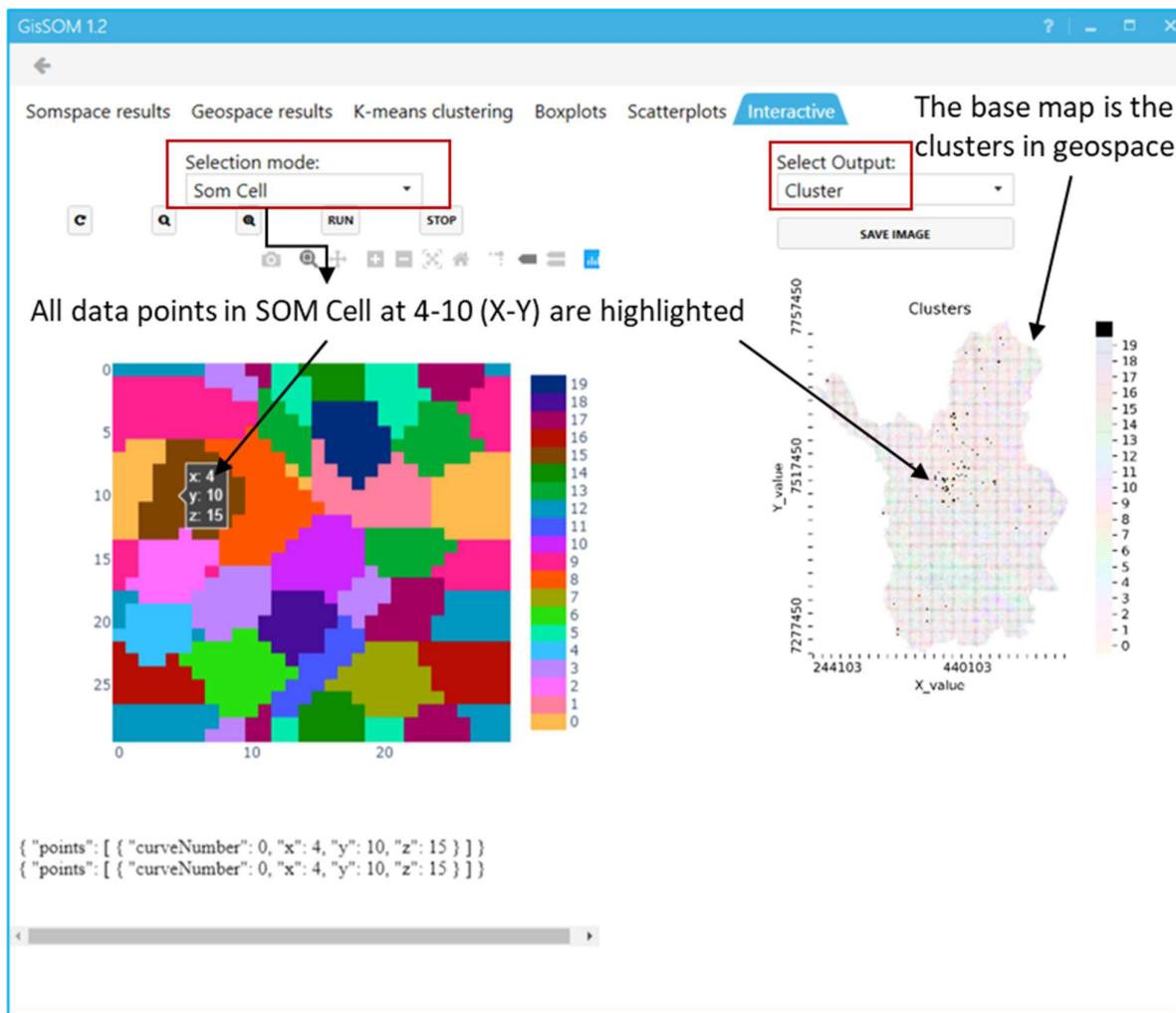


Figure 19. Interactive plot highlighting cells in geospace corresponding to a single SOM neuron.

December 5, 2022

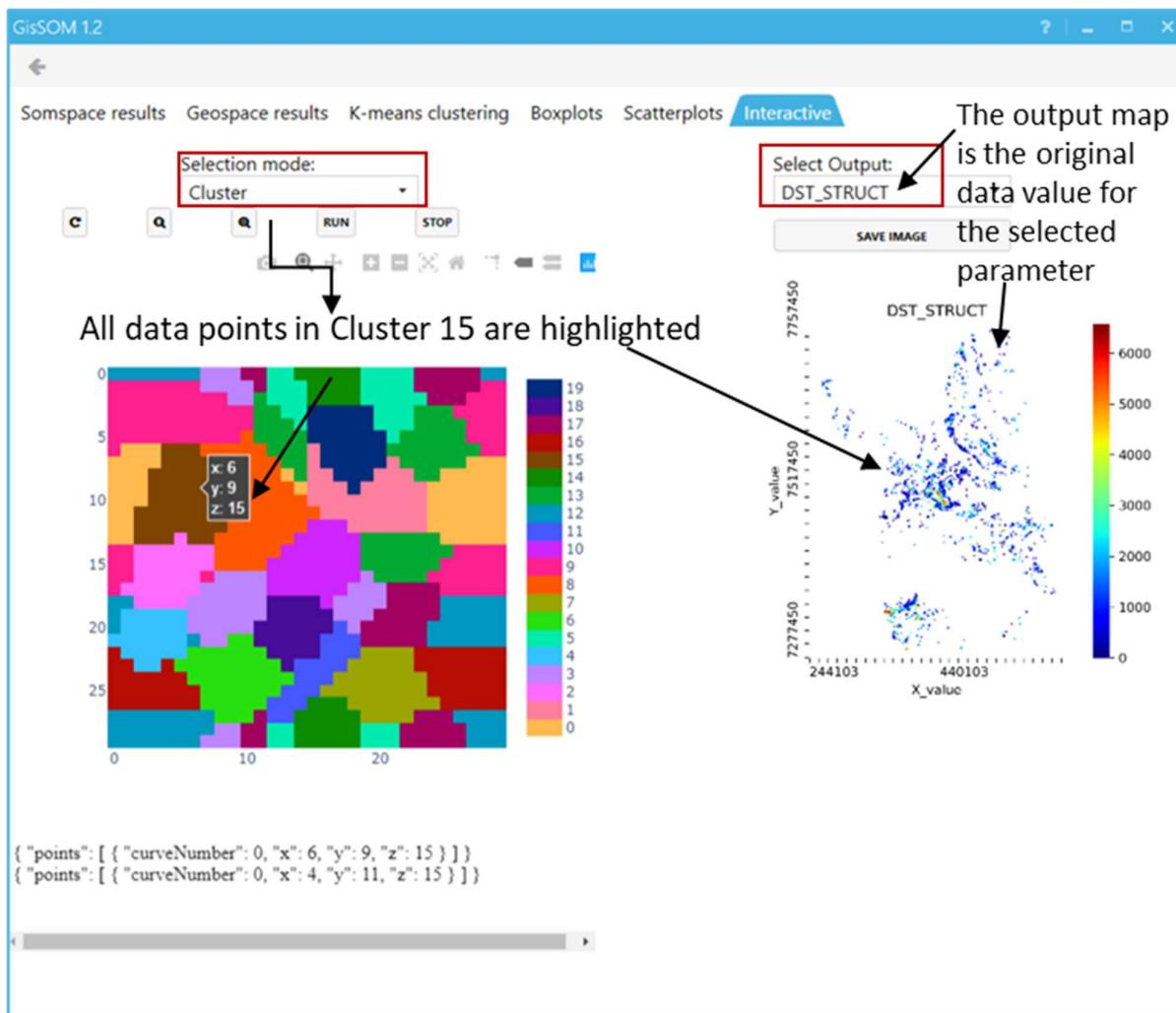


Figure 20. Selected cluster highlighting a variable of the original data.

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”-button, or if this doesn’t work, you can try stopping the plot (“Stop”-button) and running it again (“Run”-button). After the loading icon for the new plot has completed, the plot may require refreshing to display. If all else fails, you 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 “<http://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 data points on the main window’s geospace plot.

December 5, 2022

3.5 Visualizing previous results

Results from previous SOM runs can be viewed in GisSOM by selecting the “Old Results” -option from the data selection drop down menu in the data preparation view (Figure 21). Selecting a result folder (a folder prefixed “Out_” in the “GisSOM” main folder) will open the Results-window, showing the selected results. You can use the interactive plot, re-do clustering and save the new clustering scheme in the results folder of the previous SOM run. You can also run the SOM calculation again for the same input data that was used for the previous SOM run. In this case, a new results folder is generated in the “GisSOM” main folder. If the input data for the previous SOM run had been in geoTIFF format, saving the geotiff output of the new SOM run requires that the path to the original geoTIFF input files has not been changed.

It is recommended not to navigate back to the main “Input data” window unless a new project with new input data needs to be started. The original data of the previous SOM run will not be displayed here, and it cannot be re-preprocessed (Figure 22).

December 5, 2022

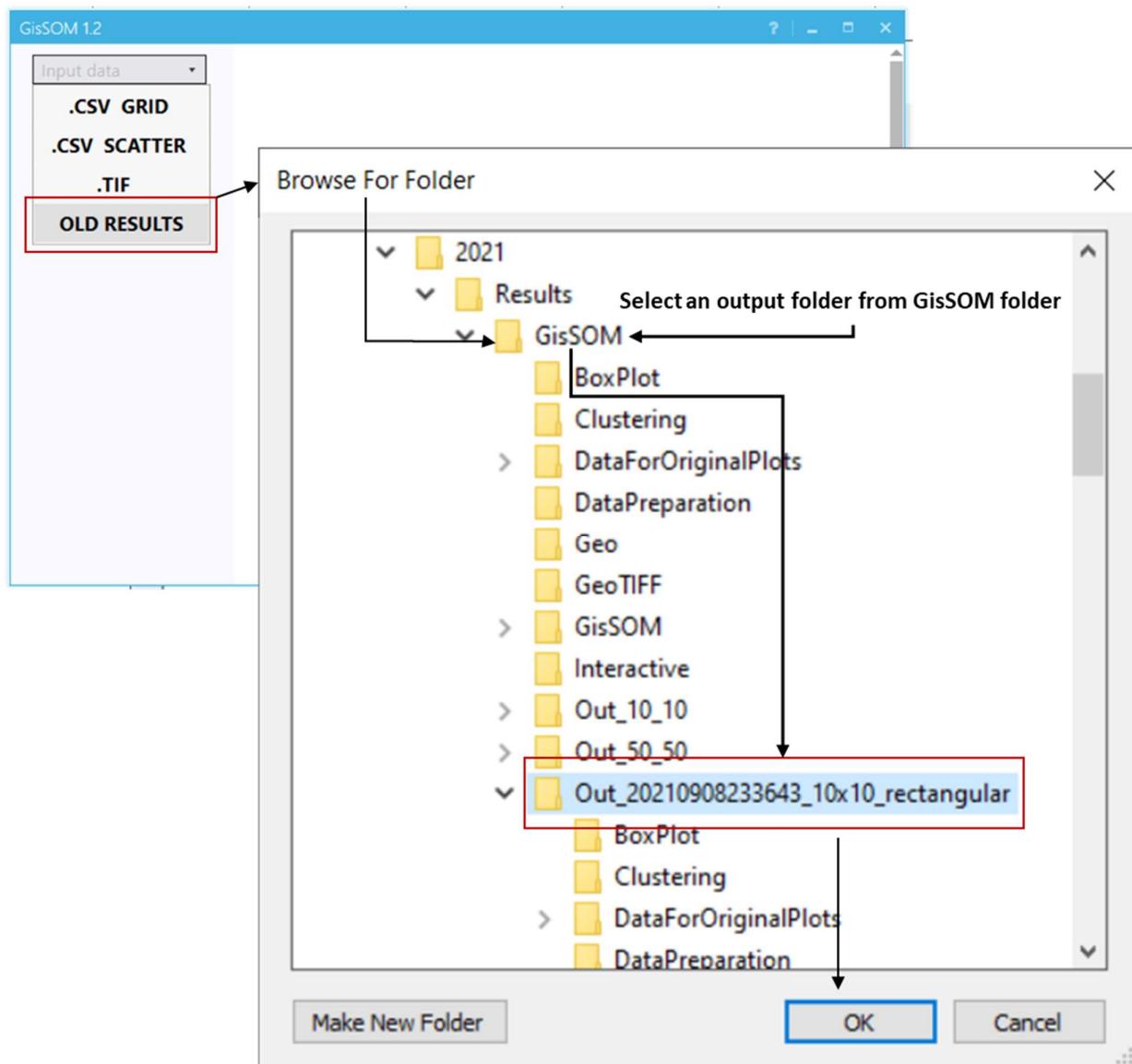


Figure 21. Visualizing previously computed results

December 5, 2022

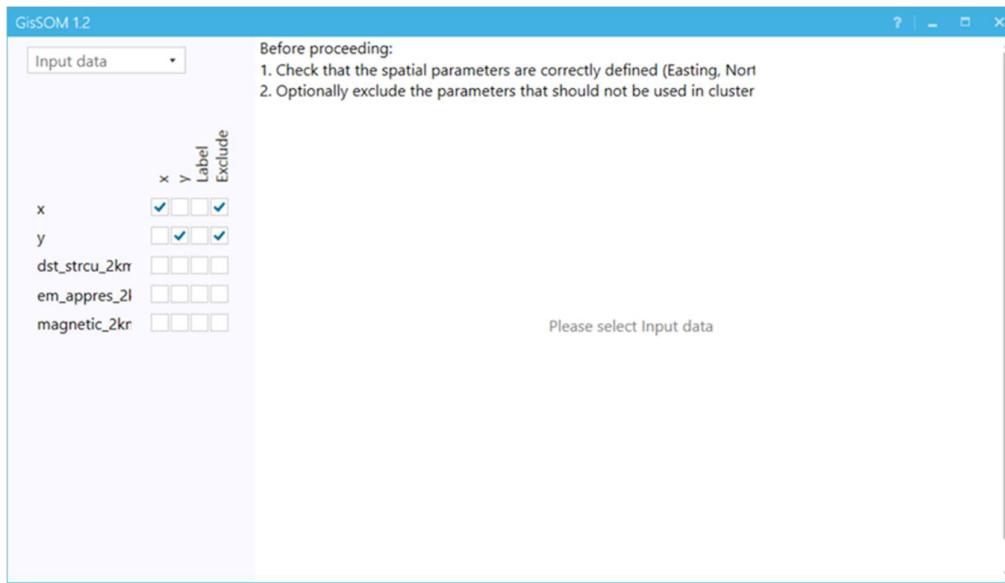


Figure 22. Loading previously computed results – The original data cannot be visualized and re-preprocessed in the Data Preparation window when using previously computed results.

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.