Assignment 3: SOM - implementation option

Evaluation Report

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# Introduction

We selected to implement the visualization group a, in particular following 5 visualizations:

* D-Matrix, P-Matrix, U\*-Matrix (U-Matrix reused from the provided template)
* qe, mqe

The implementation is available here: <https://github.com/Grave/SOM_Visualizations>

We perform a comprehensive analysis of the implemented visualizations using two benchmark datasets - 10 Gaussians and Chainlink (to be found at <http://www.ifs.tuwien.ac.at/dm/somtoolbox/datasets.html>).

For the analysis we utilize the pre-trained the SOMs, fit by Java SOMToolBox. For both datasets, the maps of size 40x20 and 100x60 were trained. Before training, the input data was normalized using the MIN\_MAX method.

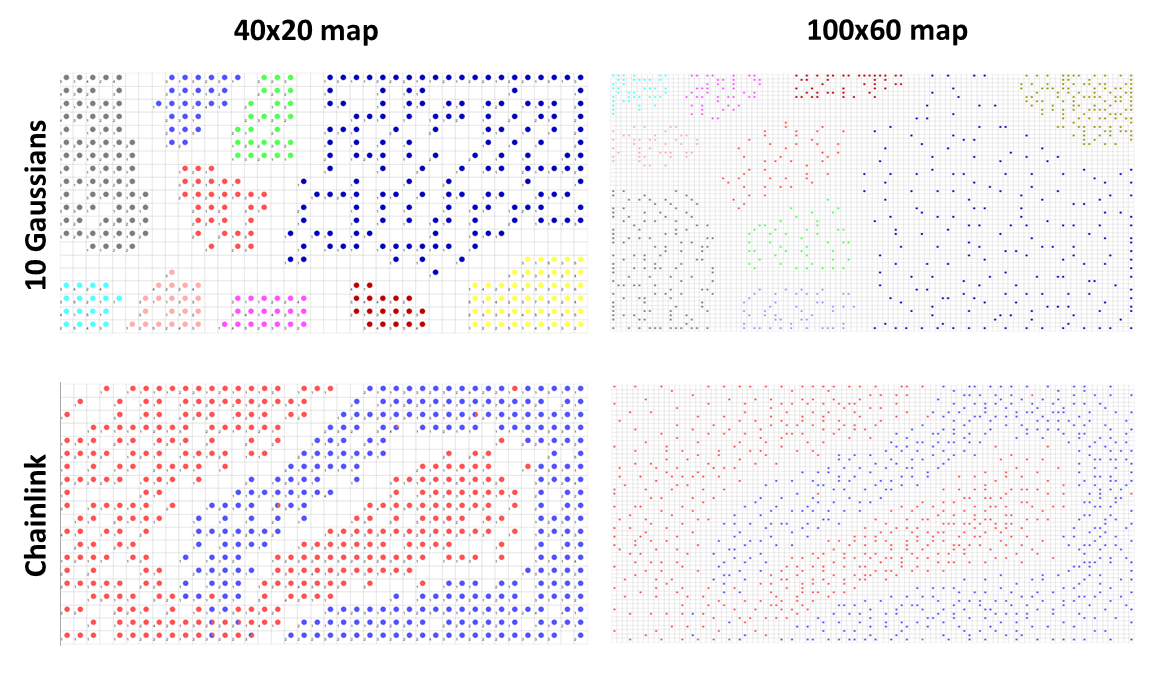
For **10 Gaussians dataset** we trained the SOMs with the following parameters:

* randomSeed=10
* numCycles=1000
* learnRate=0.7

For **Chainlink dataset** we trained the SOMs with the following parameters:

* randomSeed=10
* numCycles=1000
* learnRate=0.7
* sigma=7

We consider that these SOMs were properly trained since the respective structures (clusters or "rings") are visible, see figures below.



# Comparison against SOM ToolBox

Before starting to explore the visualizations, we validate our implementation against the visualizations produced by Java SOM ToolBox. We selected 10 clusters dataset and both small / large maps to create for the comparison all visualizations using our implementation and the ToolBox.

We use the same parameters for both approaches. For the 40x20 map we use a same radius value of 0.04 for P-Matrix and U\*-Matrix and for the 100x60 map, we use 0.05. Other visualizations have no parameters.

We see that our implementation produced very similar results to SOM ToolBox (see figures under Comparison 1 and Comparison 2). To be precise, we flipped the plots obtained with SOM ToolBox to have an easier comparison and the same positioning of the clusters. We may add that due to different color schemes some aspects of the data could be perceived better in SOM ToolBox, e.g. small clusters in D-Matrix of 40x20 SOM (dark blue regions in SOM ToolBox separated by lighter blue borders can be much easier perceived as clusters than in our visualization). However, this is an issue of color mapping and not the implementation itself, especially since we provide a possibility to choose any other matplotlib colormap instead of our default "viridis".

We did not check every single point of the visualizations, so we will not state that the results are identical, but as we did not observe any major differences, we shall consider our implementation as sufficient and will use it for the further analysis.

The comparison of visualizations for Chainlink dataset can be found in the appendix. The visualizations did not show any additional differences, so we do not include them in this comparison.

## Comparison 1 - 10 Gaussians dataset, Small SOM (40x20)

|  |  |
| --- | --- |
| **Our implementation** | **SOM ToolBox** |
| **D-Matrix** | |
|  |  |
| **P-Matrix** | |
|  |  |

|  |  |
| --- | --- |
| **U\*-Matrix** | |
|  |  |
| **Quantization Error** | |
|  |  |
| **Mean Quantization Error** | |
|  |  |

## Comparison 2 - 10 Gaussians dataset, Large SOM (100x60)

|  |  |
| --- | --- |
| **Our implementation** | **SOM ToolBox** |
| **D-Matrix** | |
|  |  |

|  |  |
| --- | --- |
| **P-Matrix** | |
|  |  |
| **U\*-Matrix** | |
|  |  |
| **Quantization Error** | |
|  |  |
| **Mean Quantization Error** | |
|  |  |

# Analysis - 10 Gaussians dataset

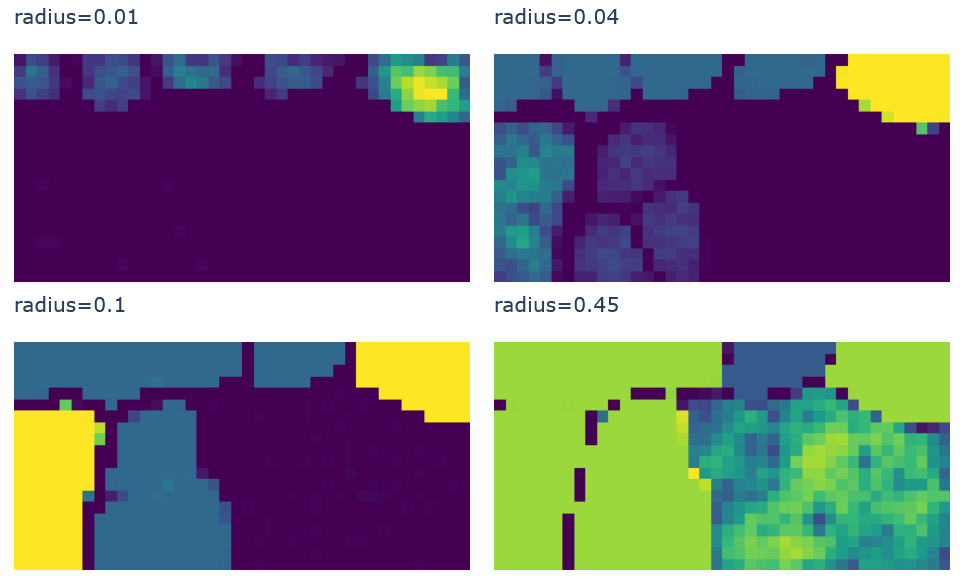
As a starting step, we analyze the effect of the parameters for P-Matrix and U\*-Matrix. Other implemented visualizations have no parameters, so we will look at them later. We know the structure of the dataset in advance (10 clusters of different sizes and densities), so with our visualizations, we aim at showing the separation between the clusters.

For a better overview of the parameters, we plot the P-Matrix for 40x20 SOM and U\*-Matrix for 100x60 (even though using a 100x60 map would be more visually appealing for both visualizations, we preferred to demonstrate the effects on the different maps). We see that in the case of the P-Matrix visualization, the radius is very important for the interpretation of the possible patterns in the data. When using a low radius, e.g. in our case 0.01, only very dense clusters can be detected. On the other hand, when the radius is selected rather higher, e.g. in our case 0.1, the clusters located close enough to each other are grouped and are perceived as one big cluster; this effect is seen even more when the radius is increased to 0.45, where some clusters could still be separated, but the borders are not so clear anymore. We select the radius 0.04 to be optimal for 40x20 SOM, as in this case the clusters are separated the best but still, only nine dense clusters can be detected and the sparse one is missed (or rather would not be considered as a cluster).

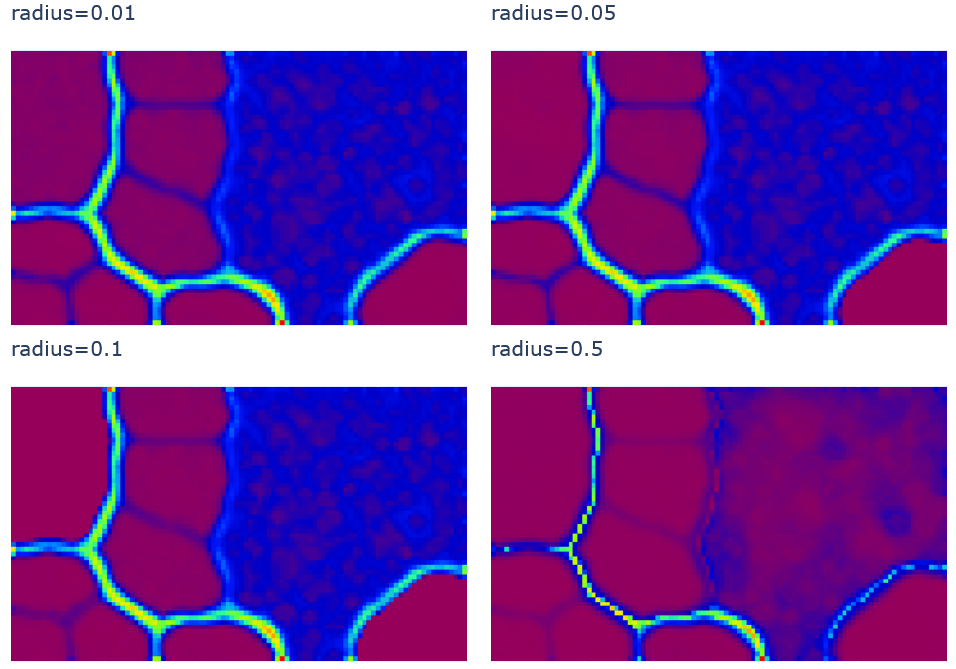
A slightly different picture is observed for U\*-Matrix. This matrix is known for its ability to separate the clusters well. Indeed, the clusters could be easily identified for different radius values - 0.01, 0.05, and 0.1 (we selected another colormap to make the borders perceptually clearer). Although for high radius values, e.g. 0.5 (recall that the initial data was MIN\_MAX scaled), the smaller clusters are not separated so well anymore, they still remain visible to some extent. For 100x60 SOM and U\*-Matrix visualization, we consider the radius 0.05 to be optimal, even though a slightly smaller or larger radius would be fine as well.

So depending on the need for interpretation of the patterns, P-Matrix with a small radius could be selected, when very "detailed" information has to be identified and missing the "big picture" or overview is not so important, or on the contrary P-Matrix with a bigger radius would be required, if a rather general view on the data is preferred (obviously, several options could and should be tried to have better insights into the data). If the goal is to identify the clusters "as they are", U\*-Matrix shall be used.

**SOM 40x20, P-Matrix**



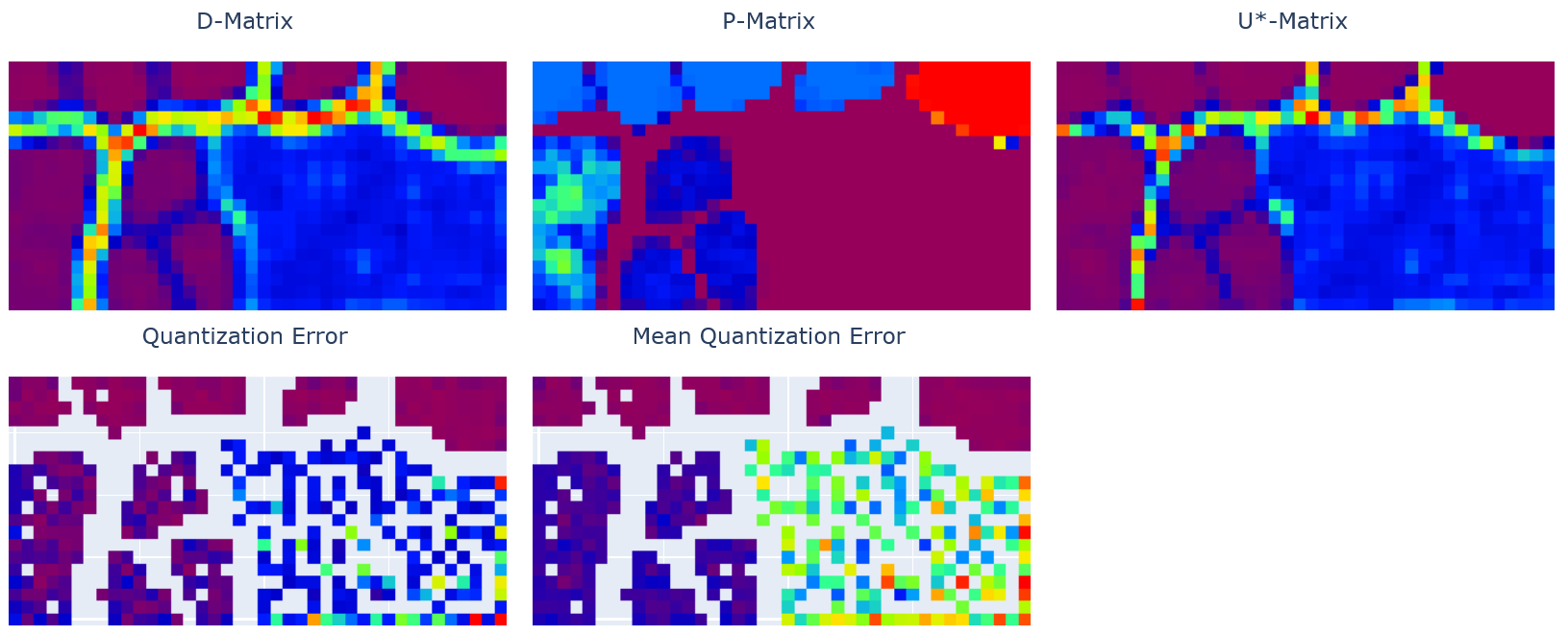
**SOM 100x60, U\*-Matrix**



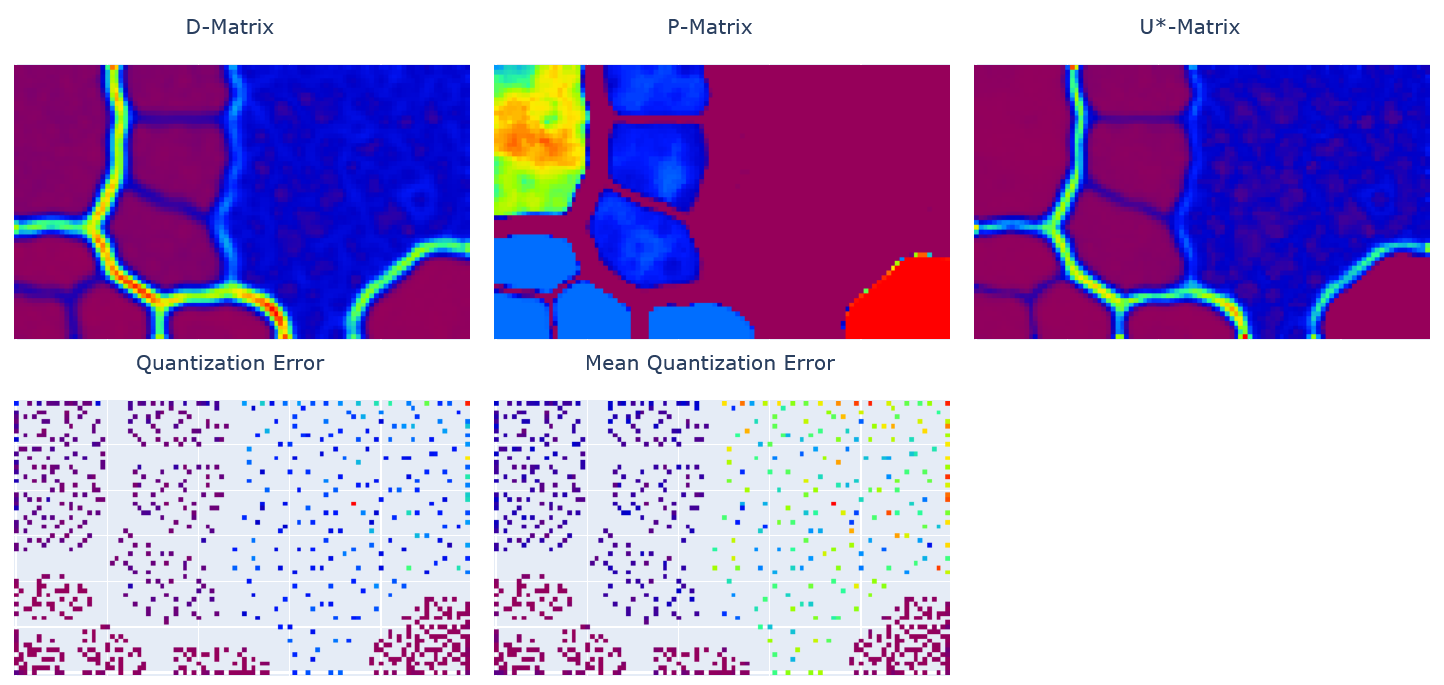
We proceed by looking at all visualizations, first for 40x20 SOM and then for 100x60. As noticed in the step before, we shall change the colormap, so that the cluster borders are visually easier perceived.

We can see that the clusters are much better separated on a 100x60 map with very clear borders in all Matrix visualizations. On the contrary, the 40x20 map performed not so well in separating smaller clusters even though they are visible. Still, without former knowledge of the structures in the data, it would be more difficult to judge clusters. Both maps failed to detect all 10 clusters and instead suggest the existence only of 9 clusters and missing the sparse cluster. When looking at the quantization error (and its mean) we see that there are higher quantization errors for units in the sparse cluster due to its spread-out nature.

**SOM 40x20**



**SOM 100x60**



# Analysis - Chainlink dataset

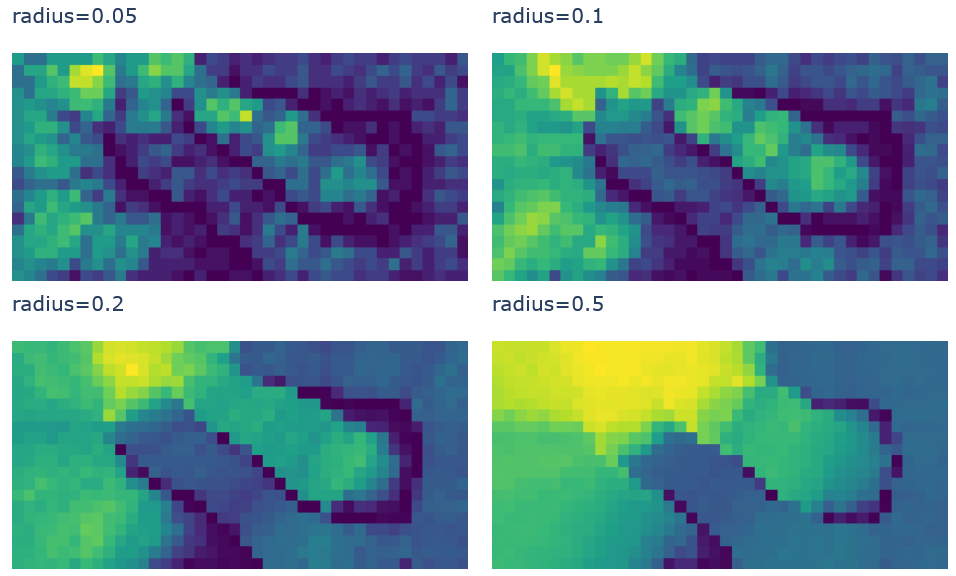
Similar to the 10 Clusters dataset, we first compare the influence of different parameters on P-Matrix and U\*-Matrix results. As the chainlink dataset represents two intertwined rings and the data is therefore not separable, we do not observe here clear structures. We will see further that these types of visualizations may be not the optimal ones for this dataset.

Still, analogically to the 10 clusters case, we observe that proper radius selection may be more informative. For example, a P-Matrix with a radius of 0.1 or 0.2 would suggest to us that there is some data separation in the initial dataset, also we see structures like rings. While in the case of a radius of 0.1 the separation seems more clear. Visualization with a radius of 0.2 emphasizes more clearly the areas where the topology violations appear. Visualizations with a radius of 0.05 or 0.5 do not represent the patterns in the data at all.

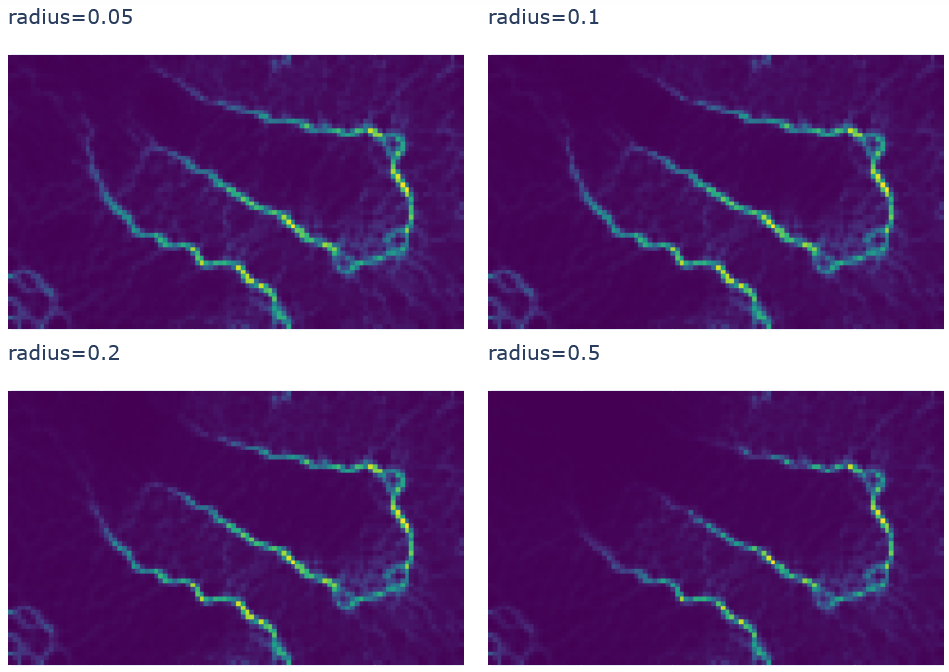
U\*-Matrix with radius values of 0.05, 0.1, and 0.2 displays very well where the data is separated. However, by looking at the visualization without former knowledge about the dataset, it would not be possible to recognize the respective structures. As identified earlier, the U\*-Matrix is not so sensitive to the selection of radius (when the considered differences are not too high).

We select an optimal radius for the P-matrix to be 0.2 and for the U\*-Matrix 0.1. However, especially in this case, this choice remains very subjective.

**SOM 40x20, P-Matrix**



**SOM 100x60, U\*-Matrix**

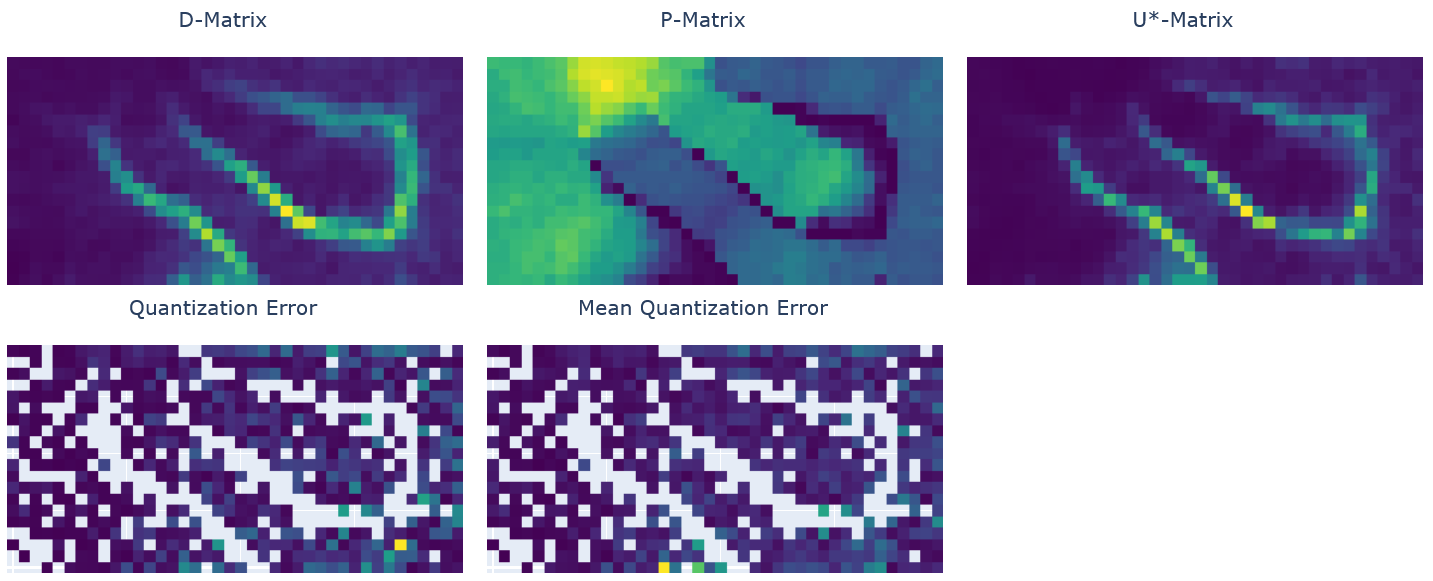


We look now at all visualizations for the Chainlink dataset and see that most visualizations are not very useful in the sense that they do not represent the pattern in the dataset well enough. An exception would be the P-Matrix for 100x60 map, where we can indeed identify the two ring structures (the regions of high density) and the position, where both rings intertwine (upper-middle with yellowish color – even higher density).

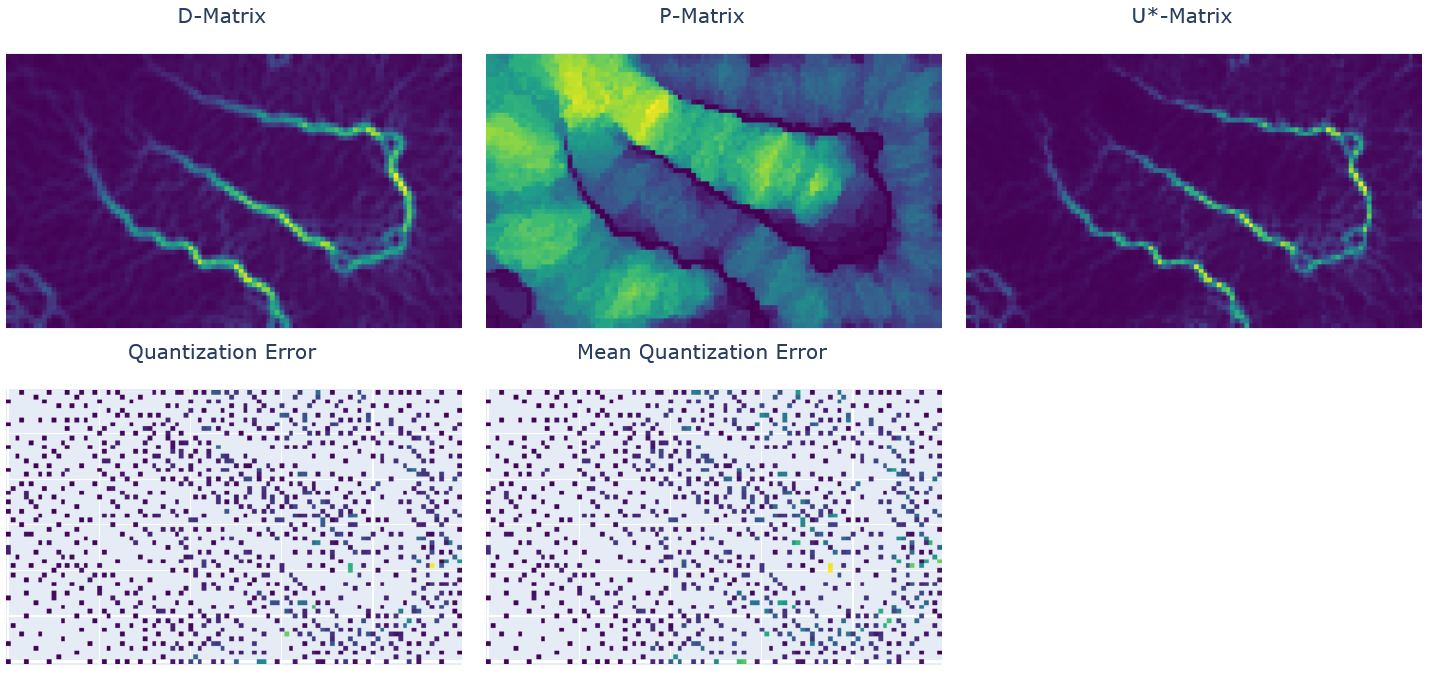
The quantization errors and mean quantization errors are distributed almost uniformly for most units of the map and therefore not so meaningful for interpretation (however from 40x20 maps we can identify the borders quite well).

At this point the selection of a proper colormap shall be emphasized again - looking at the plots produced by SOM ToolBox (see Appendix), we can confirm that P-Matrix is most representative, but the colors used in SOM ToolBox support an easier interpretation of the data.

**SOM 40x20**



**SOM 100x60**



# Conclusion

We implemented and analyzed 5 different visualizations for Self-Organizing Maps. We can conclude that for every dataset different visualization types and properly selected parameters would be more helpful to detect the structures in the original data. Moreover, a careful selection of colors is also necessary for appropriate visual representations, which would be comfortable for the human eyes to perceive. Finally, the best approach would be to try different possibilities to get a qualitative general view on the data, but still not to miss potentially important details.

# Appendix: Comparison of visualizations for Chainlink dataset

## Chainlink dataset, Small SOM (40x20)

|  |  |
| --- | --- |
| **Our implementation** | **SOM ToolBox** |
| **D-Matrix** | |
|  |  |
| **P-Matrix** | |
|  |  |
| **U\*-Matrix** | |
|  |  |
| **Quantization Error** | |
|  |  |
| **Mean Quantization Error** | |
|  |  |

## Chainlink dataset, Large SOM (100x60)

|  |  |
| --- | --- |
| **Our implementation** | **SOM ToolBox** |
| **D-Matrix** | |
|  |  |
| **P-Matrix** | |
|  |  |
| **U\*-Matrix** | |
|  |  |
| **Quantization Error** | |
|  |  |

|  |  |
| --- | --- |
| **Mean Quantization Error** | |
|  |  |