room-occupancy-vis-large

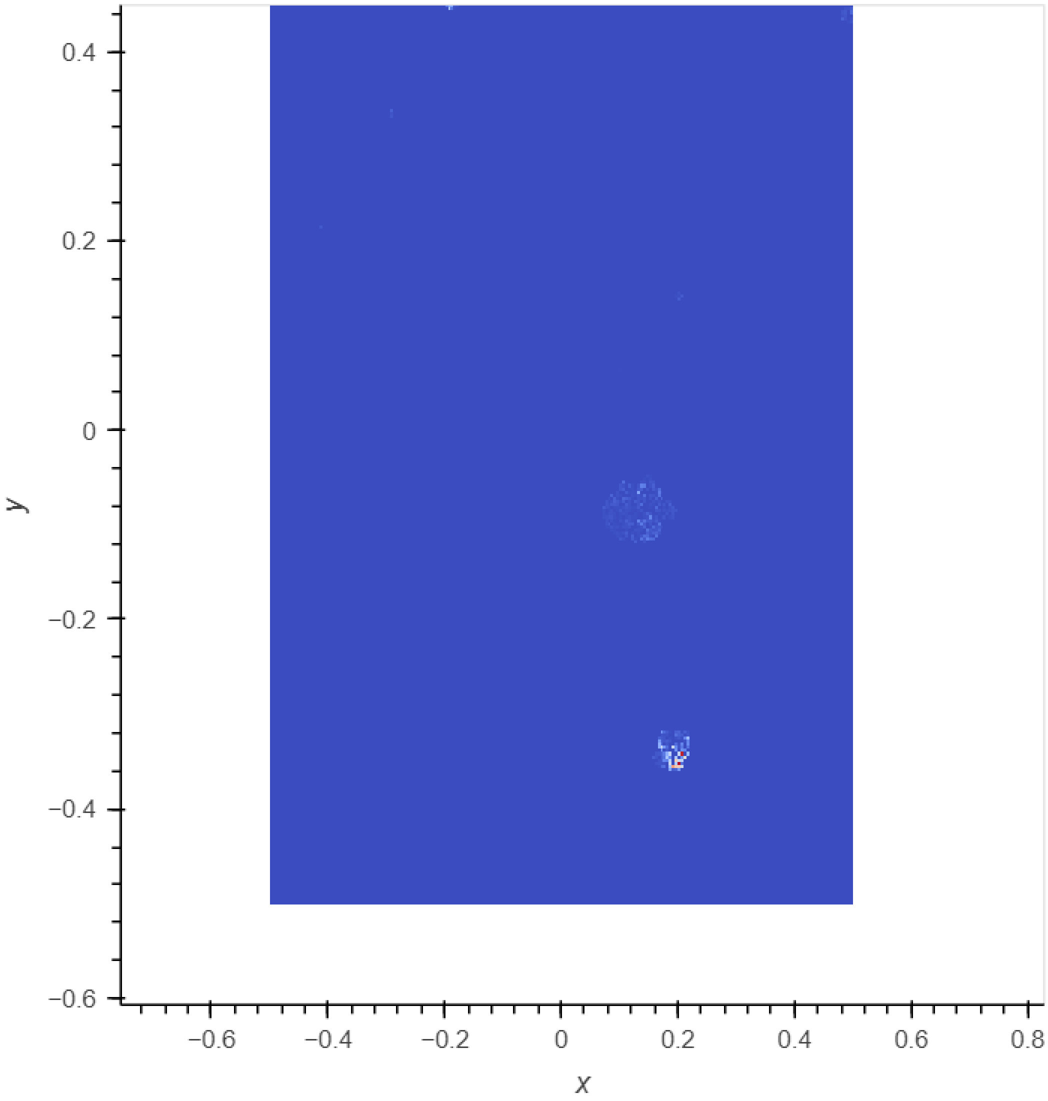
February 4, 2024

1 Visualizations large SOM - Room occupancy dataset

1.1 1. Hit histogram

The hit histogram visualizes the frequency with which neurons get hit during the training of a SOM.

We identify 3 clusters: a small one at the top, a bigger, yet sparse one in the middle and a dense one at the bottom. The neurons most frequently hit belong to the bottom cluster.



1.2 2. Smoothed data histogram

The SDH is an extension of hit histograms that maps input vectors onto n-best matching units and achieves a smoothing effect.

Compared to the visualization of the smaller dataset, the smoothing factor has less of a visible effect on this dataset. With the weighted SDH additional neurons become visible.

(Potentially Faulty) Cluster Connection Visualization

Artifacts on scrolling

1.3 3. Neighbourhood graph

Neighbourhood graphs visualize which areas of the SOM are in proximity based on the input space.

Using the radius approach, we set the radius to a value of 2.0. Same as with the smaller dataset, the connections formed are different compared to the KNN approach. Most connections are displayed between the middle and the bottom clusters with the radius setting. With the KNN Connections are predominantly formed within a cluster and some topology violations are shown between middle and top clusters.

Neighborhood Connections Radius 2

Neighborhood Connections KNN 8

Neighborhood Connections KNN 8 - Zoomed to middle cluster

Neighborhood Connections KNN 8 - Zoomed to bottom cluster

1.4 4. Sky Metaphor

Sky Metaphor is another density visualization, but maps data items on the exact position within a unit and therefore helps identify similarity between inputs within the same unit or across neighbouring units more accurately.

We observe the sky metaphor has one of the worst runtime performance for a large som. We still were able to capture a plot which is shown below and are able to inspect exact fine-grained density structures within the clusters on zooming in

Smoothing Factor 2 - Full SOM

Smoothing Factor 2 - Zoomed in

1.5 5. Activity Histogram

The Activity Histogram per data point visualizes the distance between input vector and all weight vectors.

We chose two input vectors: 0 and 816. Sample 0 represents a sample with low sensor readings in the early mornings as opposed to sample 816 from 18:16 on the same day which has high sensor readings. Both samples show cluster homogenity (gradients in distances clearly shown - instead high and low values as close neighbours) with one example - the bottom part of the middle cluster.

Smoothing Factor 2 - Full SOM

Smoothing Factor 2 - Zoomed in

1.6 6. Minimum spanning tree

The Minimum Spanning Tree visualizes related nodes on the map by connecting similar nodes with each other. The weights of the edges are computed by a distance metric between the vectors of the vertices and subsequently minimized. ADD CITATION https://www.ifs.tuwien.ac.at/~mayer/publications/pdf/may_icann10.pdf

Unfortunately for the large dataset Pysomvis was not able to produce the visualisation in a reasonable amount on time (On a High performance cloud server with 4 CPUs and 32GB RAM - directly run with python not via jupyter)

1.7 7. Cluster Connections

In this visualization technique, connecting lines are drawn between units based on threshold values.

We observe some issues with this visualisation on the large SOM. The connections are not rendered in the same way for the small SOM and we can produce some artifacts which only show on zooming or panning.

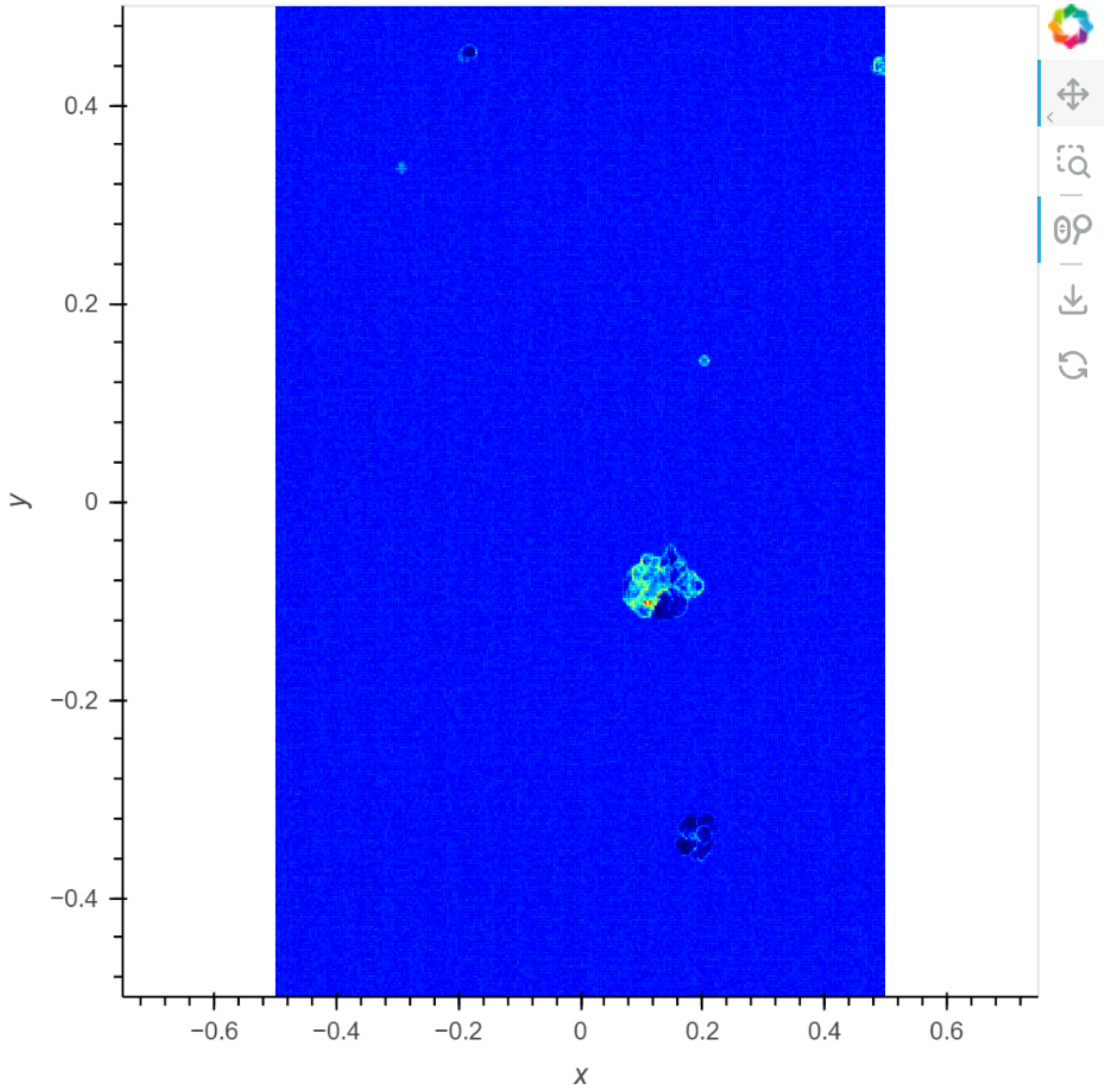
(Potentially Faulty) Cluster Connection Visualization

Artifacts on scrolling

1.8 8.U-Matrix

The U-Matrix visualization displays the distances between neurons on the SOM grid. Low values correspond to small distances between neighbouring neurons, whereas high values indicate large distances and can be used to identify cluster boundaries.

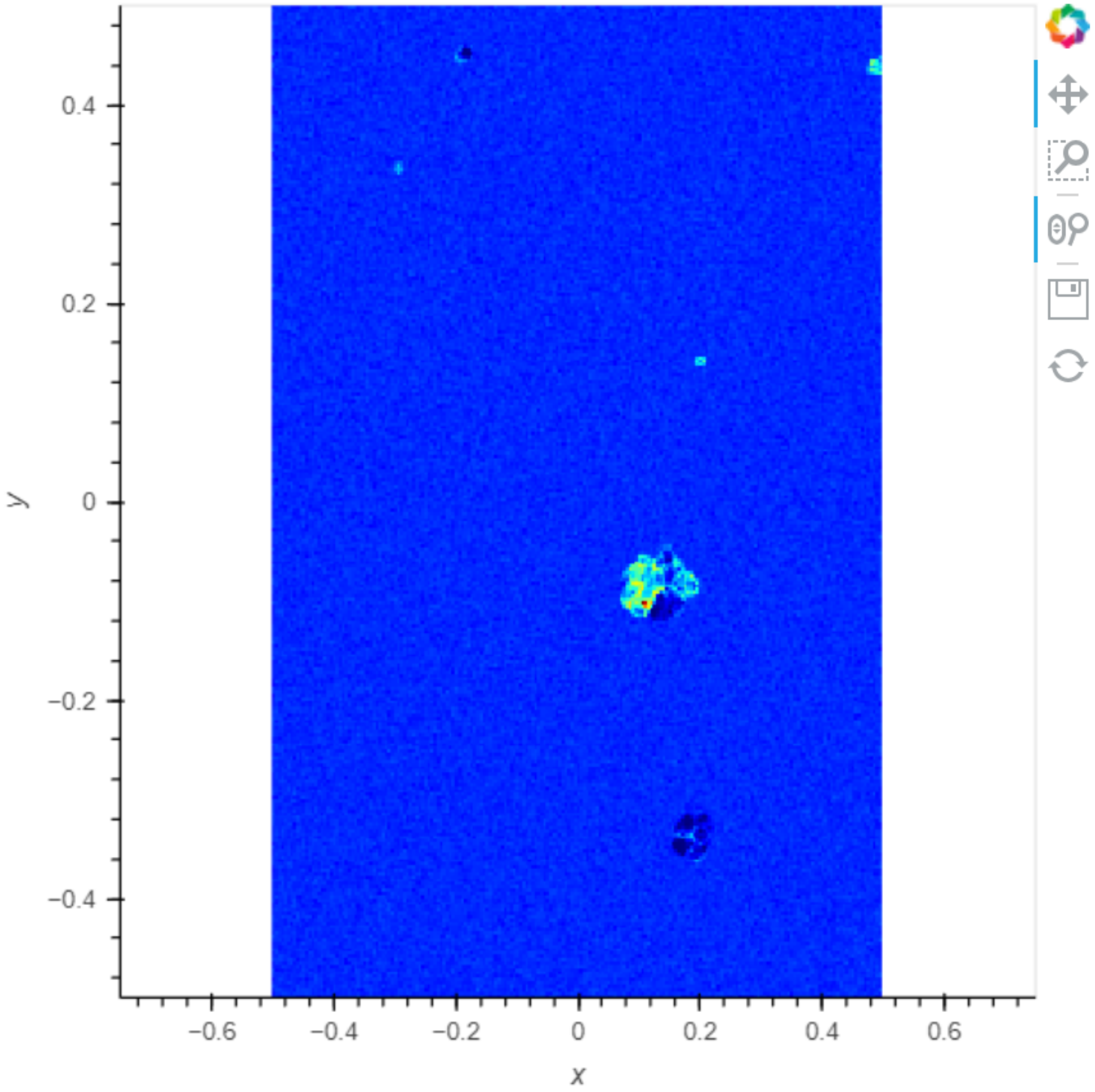
The visualization helps to discern individual cluster structures that appeared unclear in earlier visualizations. Especially the top (-0.15, -0.45) and bottom clusters (0.2, -0.3) with low sensor reading values form coherent regions(valleys) with visible cluster boundaries in the U-matrix. The middle cluster (0.15, -0.1) with high sensor readings do not form coherent, but noisy regions and unclear boundraries.



1.9 9.D-Matrix

The D-Matrix is similar to the U-Matrix, but averages the distance instead of using interpolation.

This results in a similar visualisation, but with smoother transitions between “mountains” and “valleys”. The boundaries are therefore not as clearly visible



1.10 10.P-Matrix & U*-Matrix

Unlike the U-Matrix, P-Matrix is a density and not a distance based metric. It involves estimating the empirical density at each neuron's weight vector in the feature space.

The U*-Matrix combines both distance and density information, enhancing cluster visualization by adjusting the U-Matrix with density-derived scale factors.

For the P-Matrix we calculate the optimal Percentile and Radius which results in a percentile of 43 and radius of 3.6. On increasing the percentile (and thus the radius) to 60 we can reduce the noise of the low density regions. In the P-Matrixes we observe that the supposedly 'empty' regions of the som (we see them empty in hit histogram) are shown in bright red indicating a high density region, this is something we cannot observe on the smaller SOM and have no explaination apart from an fault in the implementation on a large dataset.

In the U*-Matrix this behaviour vanishes and we observe no large structural differences to the U-Matrix indicating that Denisity information does not fundamentaly contradict the distance based

Metrics of the U-Matrix

P-Matrix with optimal values

U*-Matrix (=P-Matrix + U-Matrix) with optimal values

P-Matrix with higher percentile and radius

U*-Matrix (=P-Matrix + U-Matrix) higher percentile and radius

1.11 11. Pie chart

This visualization is for classification type datasets. The room occupancy provides the occupancy count as an integer type target, which is not suitable for this classification type visualization.

1.12 12. Chessboard

Chessboard visualization is a type of class coloring visualization, combining Voronoi Tesselation and chessboard style pixel coloring according to dominant classes.

Since the dataset is not suitable for classification type visualizations, we didn't use this visualization on the room occupancy data.

1.13 13. Component planes

The component planes visualization shows the distribution of the weights for the selected attributes (=components) across the SOM units.

We observe that the middle cluster (0.15, -0.1) represents the instances with high sensory readings for Temperature, Light, Sound, CO2 and Motion - consistently throughout those components. We can also identify the middle cluster has a smaller cluster with opposing readings (valleys) attached on its bottom side, indicating low values for Light, Sound, CO2 and Motion.

The bottom cluster (0.2, -0.3) represents instances with low sensory readings, with consistency across the components. (just a small violation within this cluster for the light component)

The last visualisation shows the time of day component. It reveals the high sensory readings occur in afternoon and evening time (middle cluster) and the low sensory readings on nights mornings (bottom cluster)

Component 0 - Temperature

Component 4 - Light

Component 8 - Sound

Component 12 - CO2

Component 14 - PIR (Motion Detection)

Component 17 - Time Of Day

1.14 14. Metro Map

MetroMap is similar to component planes, but groups weights of the selected attribute into bins. Component lines connect the centers of gravity of each bin.

When one attribute is selected with the option of 5 bins, we see how the temperature readings from the one sensor are distributed into the bins, and the component lines indicating the gradients between centers of gravity. We see the lines for the lower bins not forming a clear direction, this makes sense as the low temperature clusters are spread throughout top and bottom of the som.

If multiple attributes are selected, only one bin is visualized and. The centers of gravity for two temperature readings S3 and S4 and time of day lies in the bottom cluster so we se a clear direction/angles for the metro lines towards this cluster. For all other attributes this is not the case ad all centers of gravity, thus the metro lines are located in proximity of the central cluster.

Component 0

Component 2 (Temp S3), 3 (Temp S4), 17 (Time of Day)

All components except 2, 3, 17

1.15 15. Clustering

Clustering is a non-deterministic division of the map into regions based on the weights. There are two approaches: k-means and agglomerative clustering.

With the Kmeans Approach we receive a noisy result. This stands in contrast to the smaller som, where we do not receive such noise. The noise can be explained by neurons which are seldomly choosen as BMU (They will stay with randomly initialiyed weights). The noise increased with the cluster size.

We can remove the noise in agglomerative clustering, and can observe the inner structure of the middle (high sensor data) cluster. The bottom cluster (low sensor data) only emerges on increasing the cluster size. This indicates lower distances to the background data of the som in this cluster.

Kmeans Clustering 4 Clusters

Kmeans Clustering with 8 Clusters

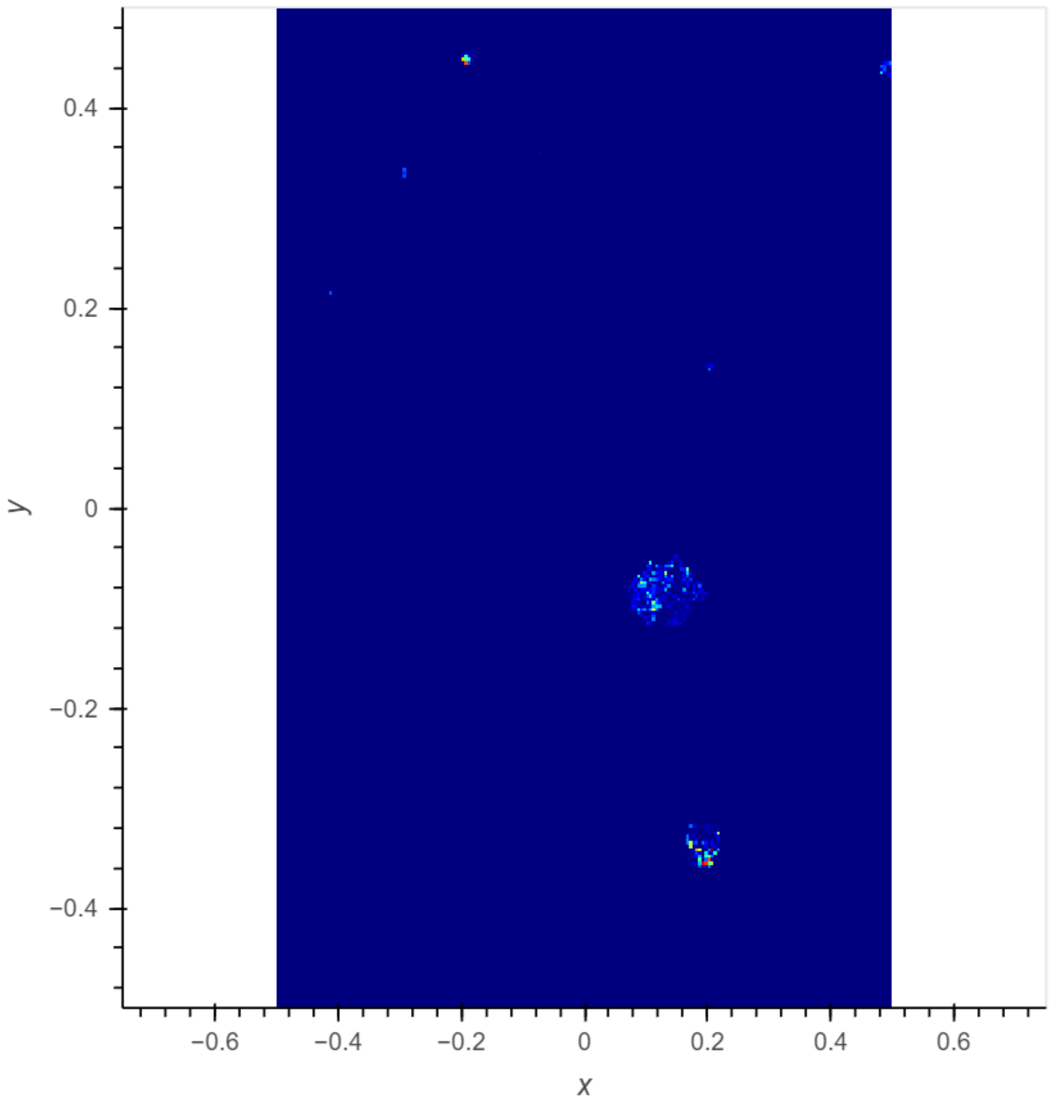
Agglomerative Clustering with 20 Clusters

Agglomerative Clustering with 50 Clusters

1.16 16. Quantization error

The visualization shows the average distance between the input vector, and it's best matching unit and serves as an indication of how well the map is trained.

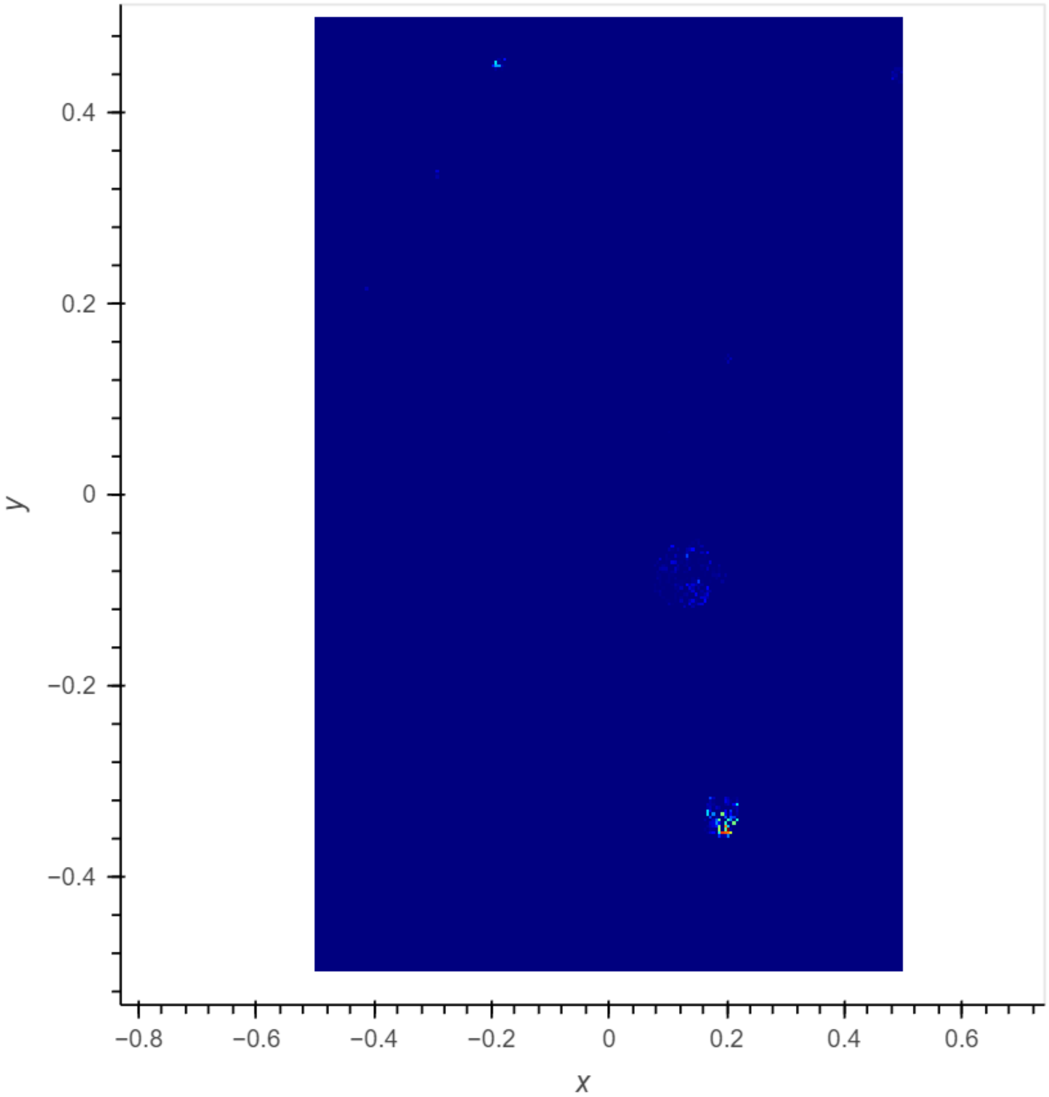
We observe a couple of single neurons with a high quant error in the top and bottom cluster, so we conclued our map is trained fairly well



1.17 17. Topographic Error

The topographic error visualizes how well the SOM preserves the topography of the input space by calculating the percentage of data samples for which the first and second BMU are not placed in adjacent units in the SOM.

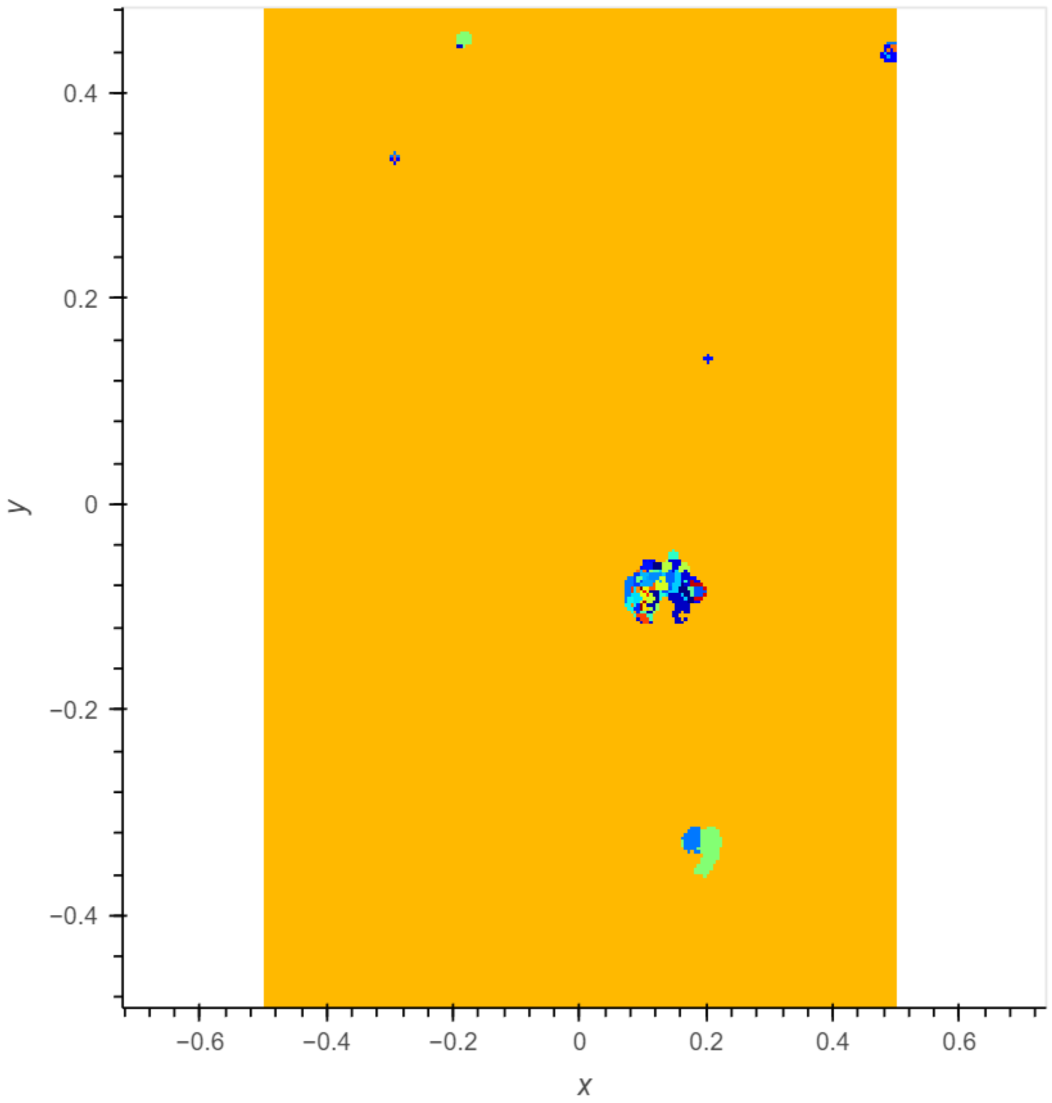
Consistent to the small som visialisation of topographic error we notice that the cluster with low sensor readings has a high amount of topographic errors, but the cluster is dense and according to the lecture this might be misleading.

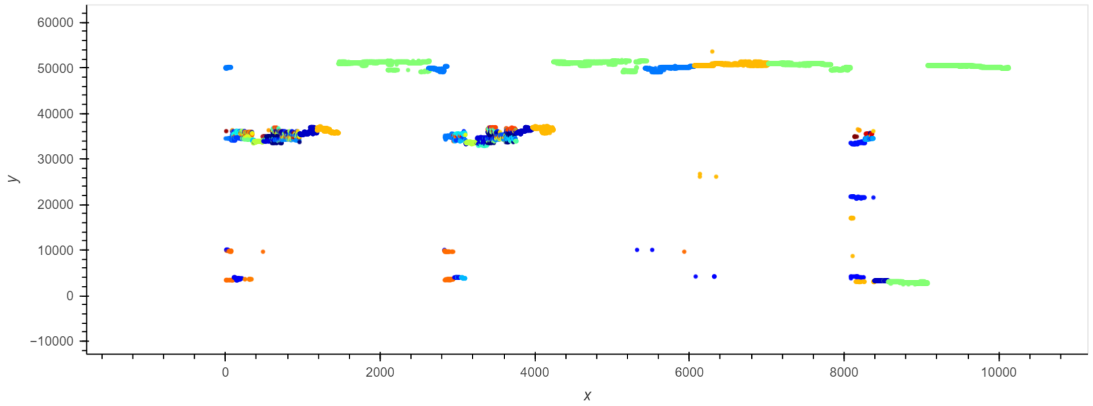


1.18 18. SOMStreamVis

The SOMStreamVis plots best matching BMU indexes over time (natural order of samples) and can provide additional information to a SOM visualization.

We used the SOMStreamVis together with the Agglomerative Clustering visualization. Colors are therefore matched between cluster and the matching BMUs over time. The dataset contains information from the 22.12. starting at 11 am until the 26.12. at 9 am, followed by a gap until the 10.01. 15:30 (at sample number 8086) and ending on the 11.01. at 9am. SOMStreamVis reflects patterns in the readings, data that belongs to the cluster with low sensor readings match samples from the nights and the high sensor readings are found in samples from afternoons and evenings. The 25.12. is an outlier in the sense that there is no match with the high sensor reading cluster in the afternoon/evening, indicating low or now occupancy at this day.





1.19 18. Intrinsic distance

Intrinsic distance visualization combines topographic and quantization error visualizations.

Due to performance issues or a bug, we weren't able to display the visualization. According to the logging we have built into it, the calculate function never terminated.

1.20 19. Mnemonic SOM

Due to performance issues on this dataset, we were unable to render this visualization. The logging we additionally implemented indicated that the calculate function never terminated.