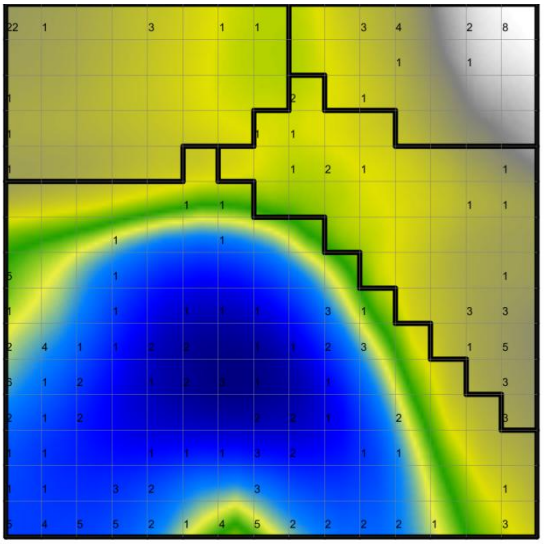
Selbstorganisierende Systeme – 3.Übung

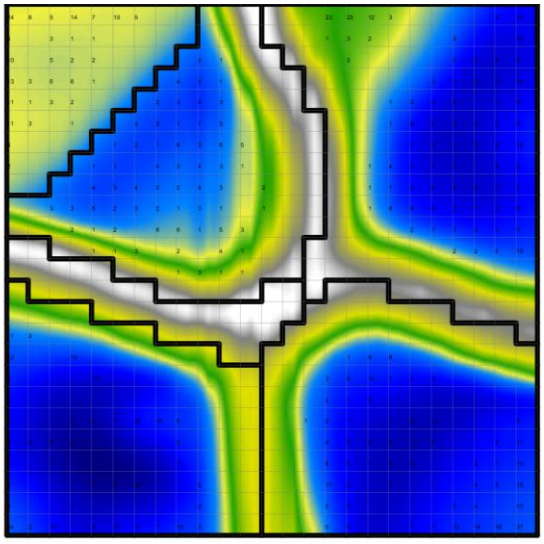
Christian Gruber, 0625102 und Johannes Reiter, 0625101

1. Clustering

Self-Organising Maps (SOM) sind eine bekannte Methode des Data Mining und werden dabei auch für das Clustering und die Charakterisierung von den zugrundeliegenden Daten eingesetzt. SOMs gehören zu den neuralen Netzen, die auf unüberwachtes Lernen basieren. Um die Daten dann auch richtig interpretieren zu können, werden verschiedenste Arten von Visualisierungstechniken verwendet. Wir haben die zwei Datasets „generic“ und „artificial“ zu untersuchen, wobei wir für das zweitere eine größere Map und wesentlich mehr Runs für das Training verwendet haben.
Um die Verteilung von Daten zu analysieren, können z.B. Hit Histogramme gewählt werden. Von diesen haben wir eine etwas fortgeschrittene Methode für das Clustering verwendet.



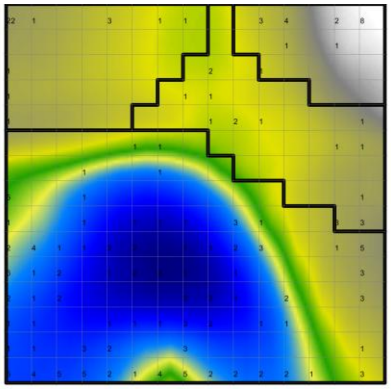
Smoothed Data Histogram mit Map Clustering k-means (Generic-Dataset)



Smoothed Data Histogram mit Map Clustering k-means (Artificial-Dataset)

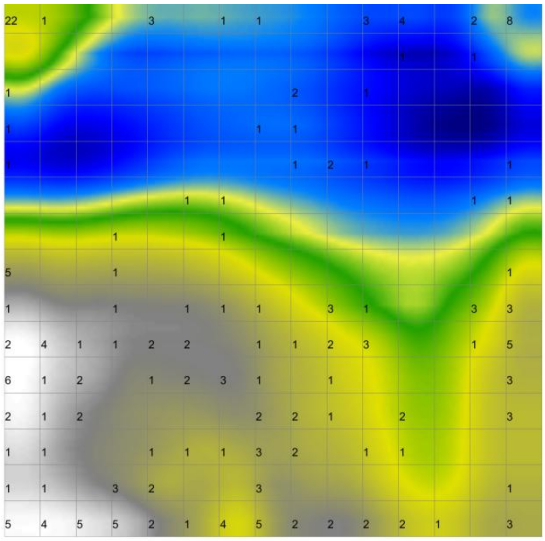
Mit dem Map-Clustering k-means und der dahinter liegenden Visualisierung des „Smoothed Data Histogram“ würden wir die entstandene Map des 1.Datasets in vier Cluster teilen. Für das 2.Dataset

hätten wir sechs „Hauptcluster“ identifiziert. Diese Methode ist nicht deterministisch, d.h. dass das Ergebnis bei gleichen Eingabedaten unterschiedlich sein kann. Ist nun ein Ergebnis bzw. dessen Cluster reproduzierbar in mehreren Runs, so kann man solche Cluster als stabil bezeichnen. Aus diesem Grund folgt hier ein weiterer Screenshot des Map-Clustering k-means für das „Generic“-Dataset, der zeigt, dass sich wirklich Grenzen leicht verschieben können, aber die Cluster grundsätzlich erhalten bleiben und deshalb in beiden Datasets stabil sind.

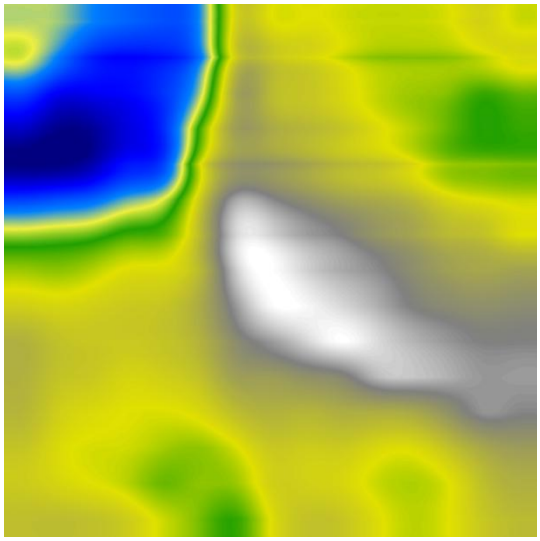


Weitere Generierung: Smoothed Data Histogram mit Map Clustering k-means (Generic-Dataset)

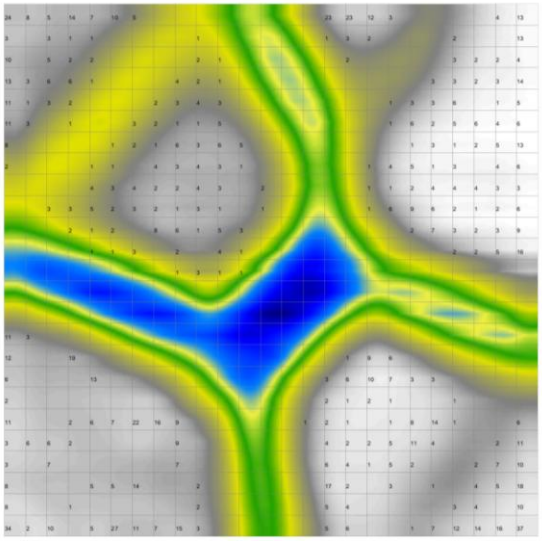
Zur Untermauerung dieser ersten Interpretation haben wir natürlich auch noch andere Visualisierung verwendet. Da eine D-Matrix und auch eine U-Matrix gut zum Veranschaulichen von Clusterstrukturen geeignet sind, haben wir diese dazu verwendet. Die D-Matrix zeigt die durchschnittliche Distanz einer Unit zu den Nachbarn. Bei der U-Matrix wird dies für alle Nachbarn berechnet, was eine feinere Auflösung zur Folge hat. Felder mit „hohen“ Werten in der U-Matrix teilen somit unterschiedliche Regionen der Input-Daten. In der U*-Matrix werden nun die Dichteinformationen aus einer P-Matrix mit den Distanzinformationen aus einer U-Matrix kombiniert. Mit dieser Weiterentwicklung, die kompatibel zum hierarchischen Clustering von Daten ist, können Cluster nun noch besser erkannt werden. In unserem Beispiel haben wir durch die U*-Matrix eigentlich keine neuen Cluster entdeckt, es wurden somit die bisher interpretierten Clusterstrukturen bestätigt.



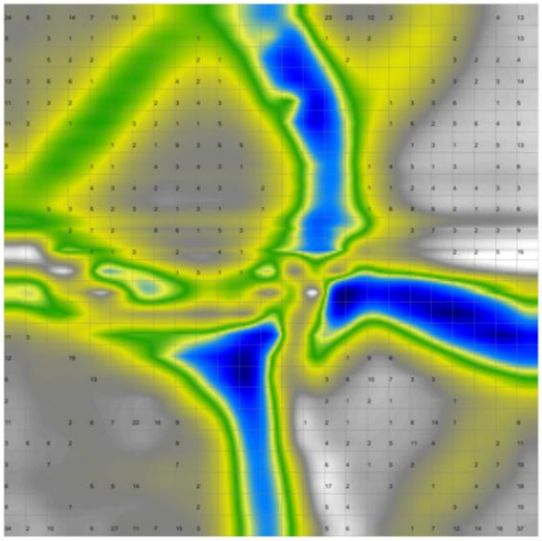
D-Matrix (Generic-Dataset)



U*-Matrix (Generic -Dataset)

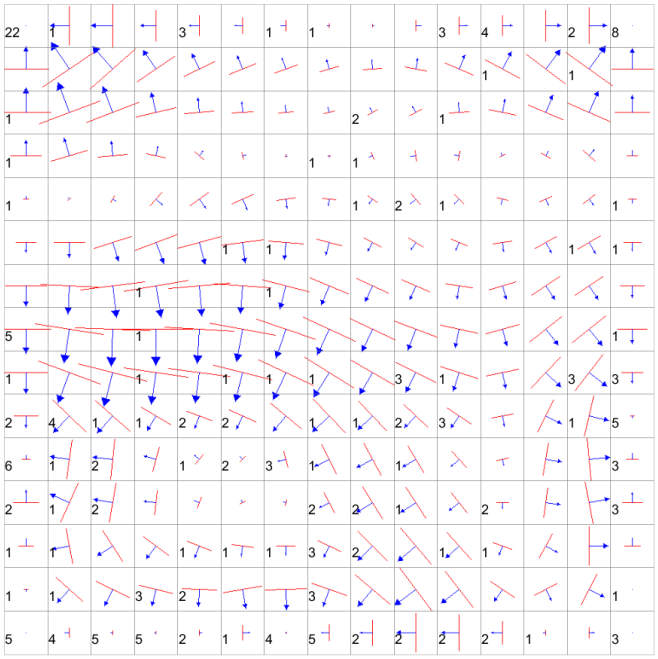


D-Matrix (Artifical-Dataset)

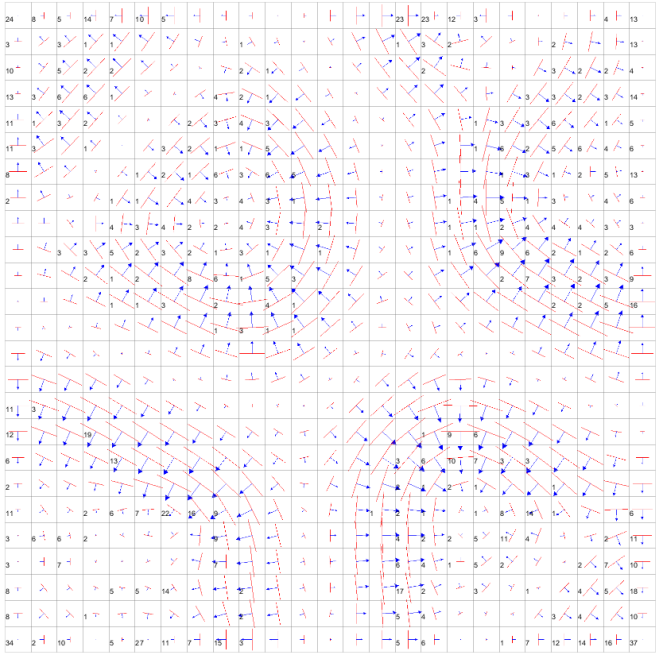


U*-Matrix (Artifical-Dataset)

Weiters haben wir uns auch noch die Flow & Borderlines Visualisierung betreffend dem Clustering angesehen. Diese Visualisierung zeigt die gefunden Clusterstrukturen wirklich sehr schön auf. Vor allem die schon zuvor durch das SDH beschriebenen vier bzw. sechs „Hauptcluster“ können hier sehr gut erkannt werden. Aber auch Subcluster, auf die wir anschließend noch ein bisschen genauer eingehen, werden durch diese Visualisierung für das 1.Datasets schon aufgezeigt, wobei es beim Artifical-Dataset eher schwierig ist, da es anscheinend keine wirklichen Subcluster gibt.

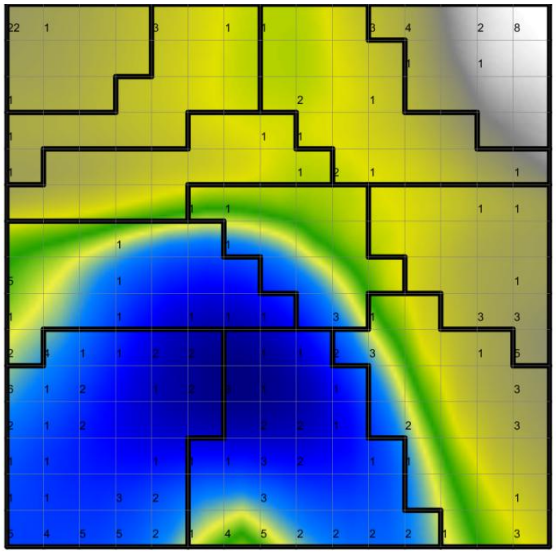


Flow & Borderlines (Generic-Dataset)



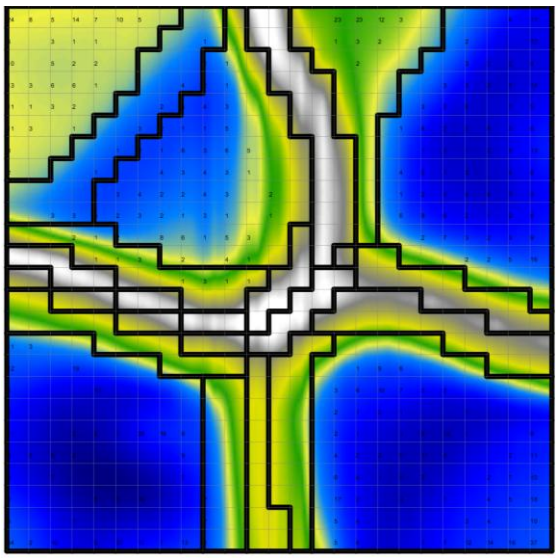
Flow & Borderlines (Artifical-Dataset)

In der folgenden Abbildung haben wir versucht die Subcluster für das 1.Dataset herauszuarbeiten. Wir haben die vier Cluster nochmals in elf Subcluster unterteilt. Das linke obere Cluster besteht aus 3 Sub-Cluster, das rechte obere aus 2, das darunter liegende aus 3 und das große Cluster links unter wieder aus 3 Sub-Cluster. Beim Subclustering haben wir uns natürlich auch wieder auf die schon zuvor verwendeten Visualisierungen gestützt.



Smoothed Data Histogram mit Map Sub-Clustering k-means (Generic-Dataset)

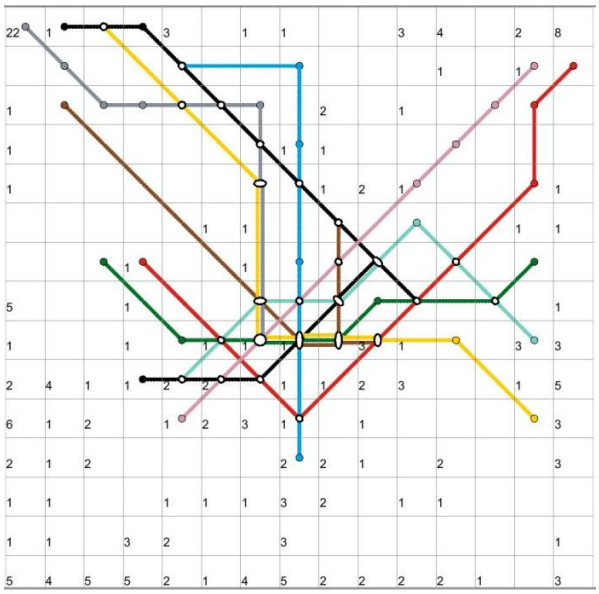
Für das 2.Dataset haben wir uns mit dem Subclustering etwas schwerer getan. Geht man von der Flow Visualisierung und dem Hit Histogramm aus, würde man 1-3 Subcluster erkennen. Mit k-means jedoch werden zuerst die freien Räume in unterschiedliche Subcluster geteilt, bevor die von uns interpretieren Subcluster eingeteilt werden.

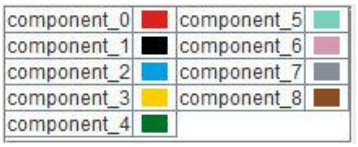


Smoothed Data Histogram mit Map Sub-Clustering k-means (Artificial-Dataset)

2. Attribute

Um die verschiedenen Komponenten und deren Beziehung untereinander analysieren zu können, verwenden wir die sogenannten Metro-Maps, in denen verschiedenste Informationen abstrahiert und vereinfacht in einer Visualisierung dargestellt werden, die wesentlich besser zu interpretieren ist. Metro-Maps helfen auch den Einfluss einzelner Variablen oder Komponenten auf die Cluster darzustellen. Die einzelnen Komponenten werden hier durch Component-Lines dargestellt. Die Component-Lines verbinden die Gebiete der Component-Planes vom kleinsten bis zum größten Komponentenwert.



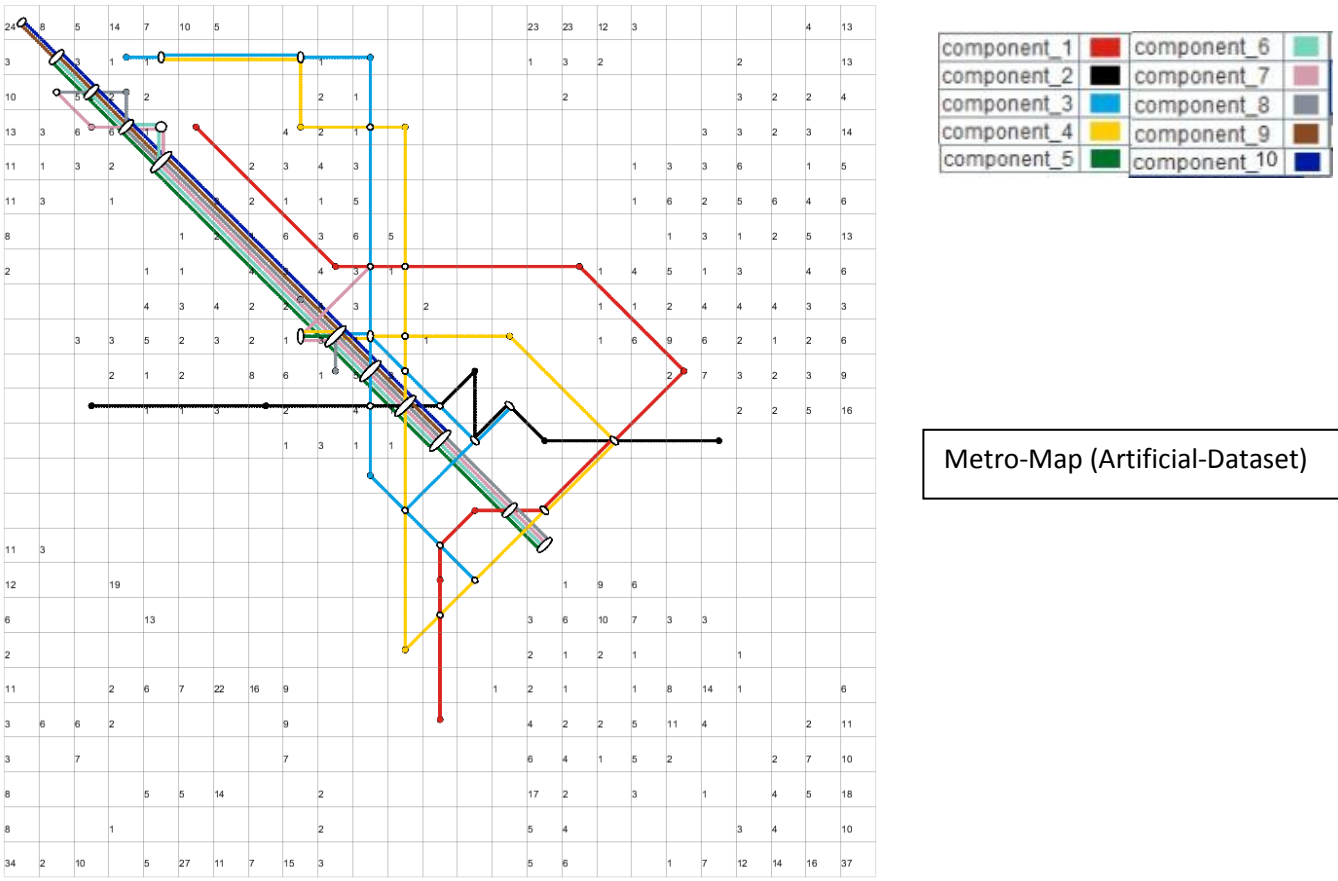


Metro-Map (Generic-Dataset)

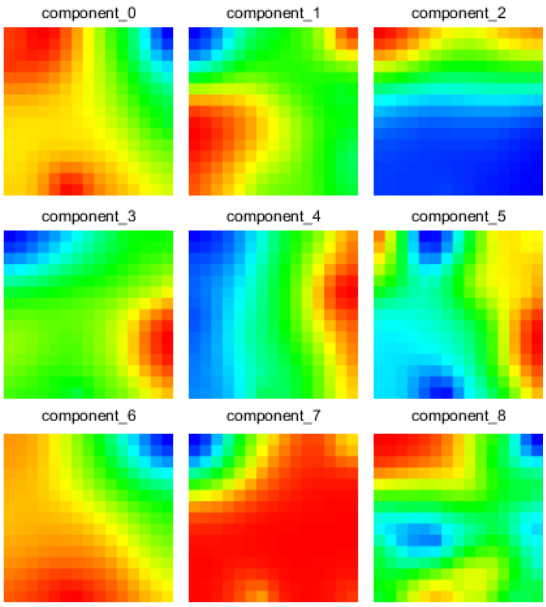
In dieser Metro-Map des 1.Dataset sehen wir neun Component-Lines. Das bedeutet in unseren Daten haben wir neun verschiedene Attribute. Die Component-Lines wurden auch mit Component-Planes, die in der unteren Abbildung zu sehen sind verglichen.

In der Metro-Map ist zur sehen dass die Komponenten 1, 2, 3, 7 und 8 den linken oberen Cluster beeinflussen. Die Komponenten 0 und 7 den rechten oberen, Komponenten 3, 4 und 5 den rechten unteren Cluster und der linke untere größere Cluster wird von allen beeinflusst. Weiters kann durch die Metro-Map auf die Korrelation der Attribute geschlossen werden. Bei uns korrelieren die Attribute 1, 2, 3, 7 und 8 aber auch 0, 4 und 6 relativ stark.

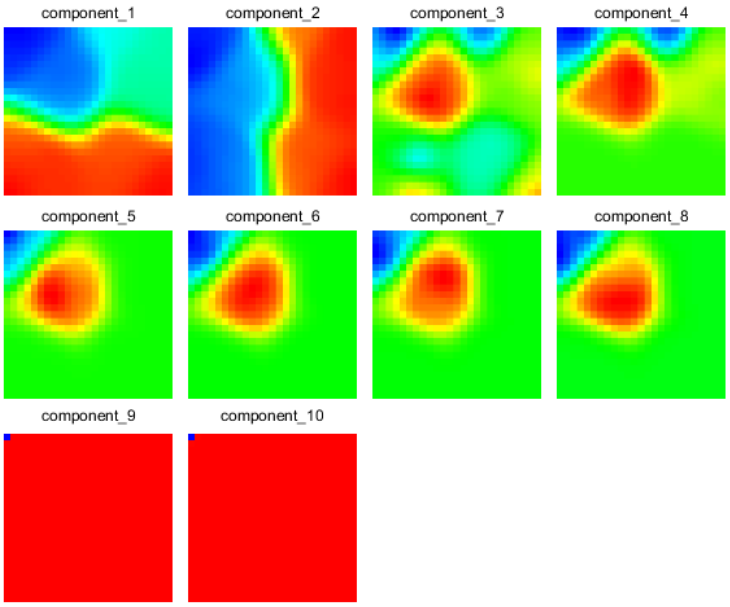
Für das 2.Dataset werden durch die Metro-Map zehn Komponenten dargestellt. Sehr interessant ist der Einfluss der verschiedenen Attribute, da fast alle vor allem die beiden Cluster links oben beeinflussen. Außerdem korreliert eine Vielzahl von Attributen (5, 6, 7, 8, 9 und 10) wirklich sehr stark. Eigentlich bleiben dann zusammengefasst nur fünf Attribute übrig, die dann wirklich eine Unterscheidung möglich machen. Eine Korrelation gibt es auch zwischen den Attributen 3 und 4.



Metro-Map (Artificial-Dataset)



Component-Plane der einzelnen Komponenten (Generic-Dataset)



Component-Plane der einzelnen Komponenten (Artificial-Dataset)

Die Frage um welche Daten es sich hierbei handelt ist natürlich sehr schwer zu beantworten. Wie haben uns gedachtet, dass es sich beim 1.Dataset um die Klassifikation von Tieren handeln könnte, die wir auch in der Vorlesung behandelt haben. Die Anzahl der „Hauptcluster“ und „Subcluster“ könnte in diesem Beispiel ca. passen, die Attribute wären dann die unterschiedlichen Eigenschaften der Tiere. Beim 2.Dataset haben wir an die Klassifikation von Autos, da hier eine Menge von „Hauptclustern“ und eher weniger „Subcluster“ entstehen könnten.