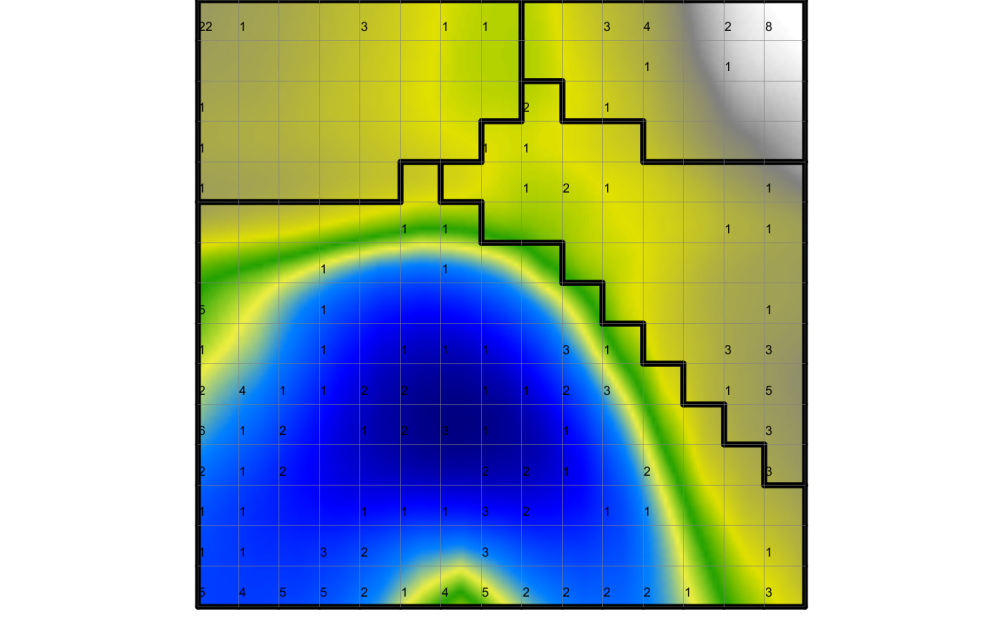
Selbstorganisierende Systeme – 3.Übung

**Christian Gruber, 0625102 und Johannes Reiter, 0625101**

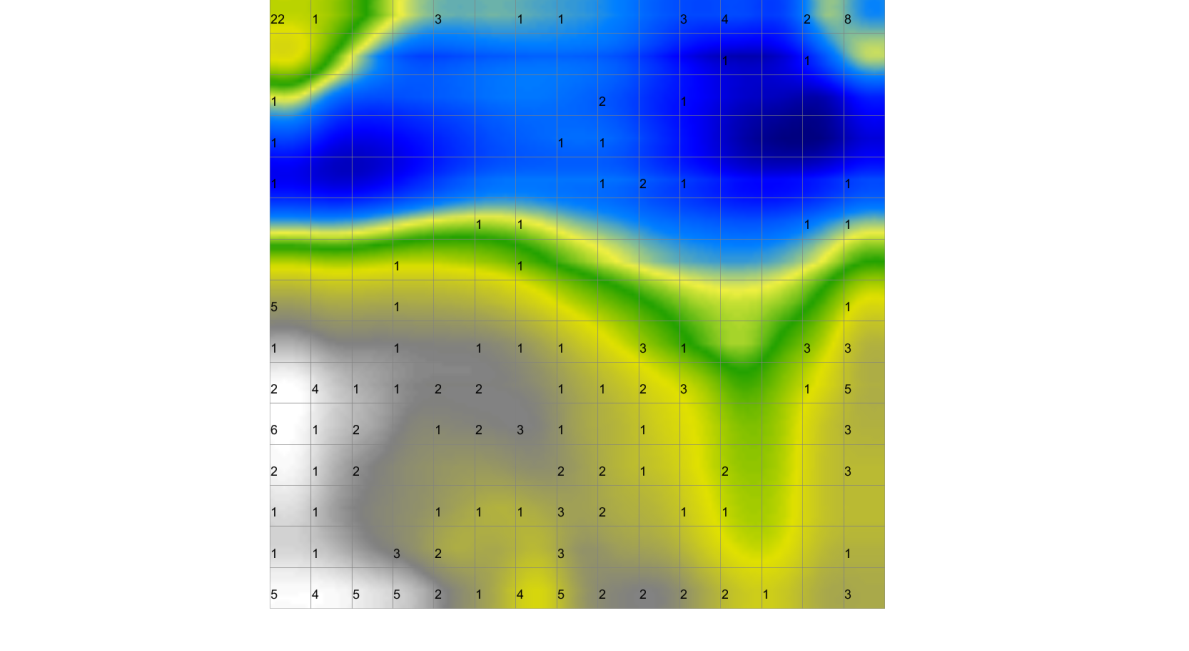
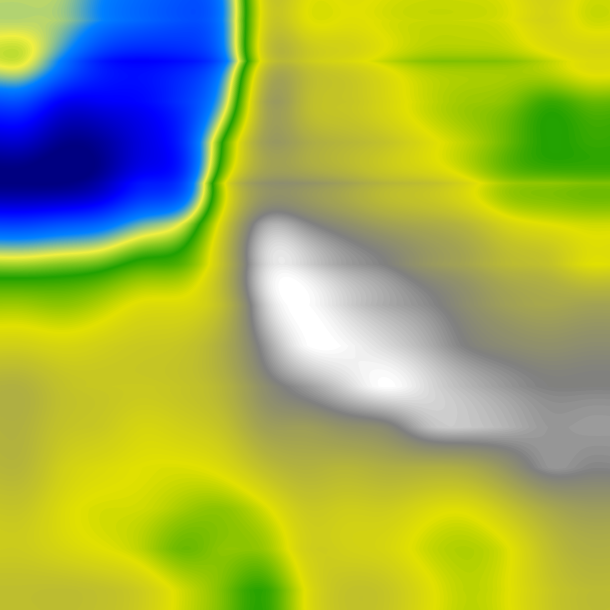
# Clustering

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. Um die Verteilung von Daten zu analysieren, können z.B. Hit Histogramme gewählt werden. Von diesen haben wir eine fortgeschrittene Methode für das Clustering verwendet.



Smoothed Data Histogram mit Map Clustering k-means

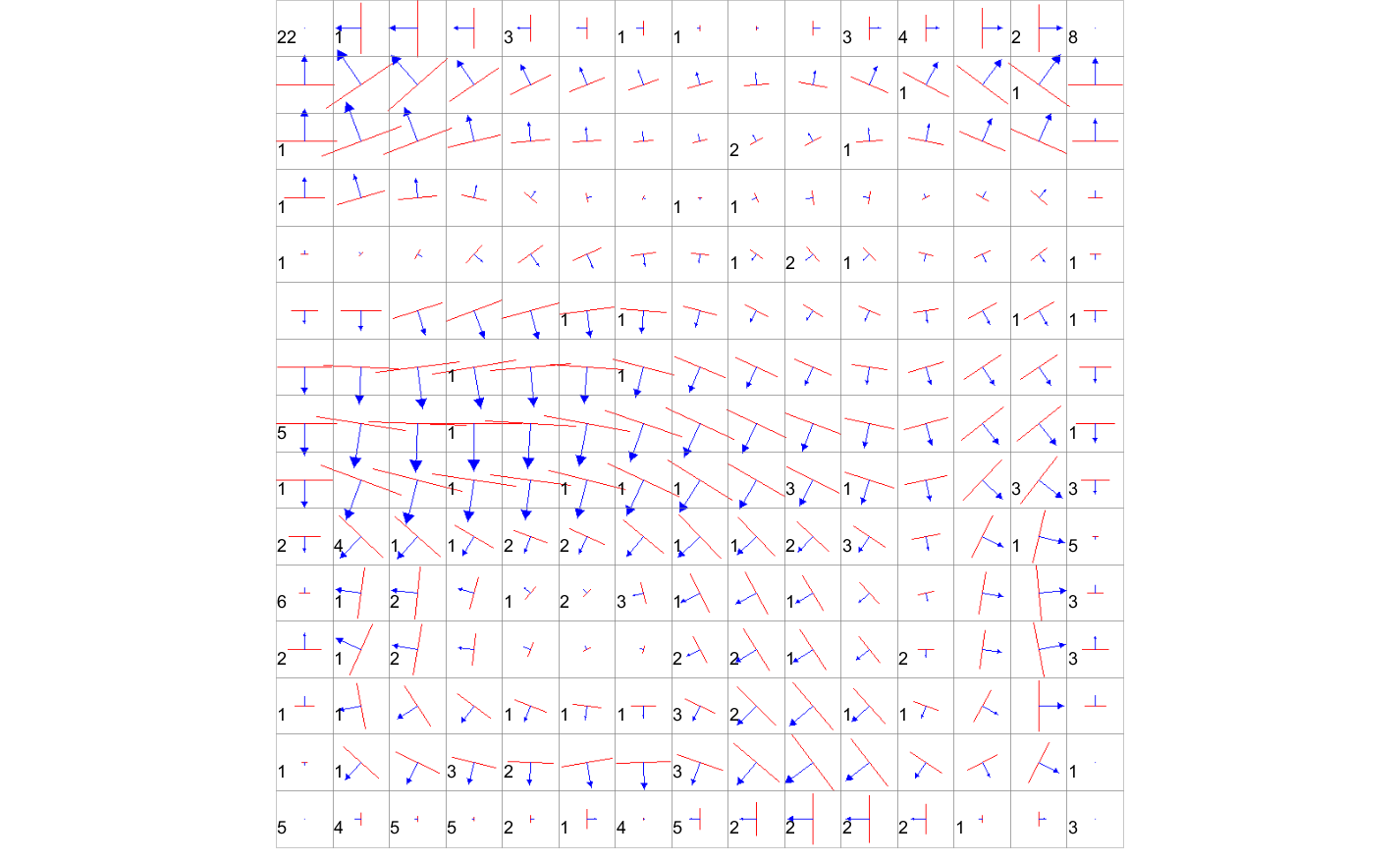
Mit dem Map-Clustering k-means und der dahinter liegenden Visualisierung des „Smoothed Data Histogram“ würden wir die entstandene Map in vier Cluster teilen. Zur Untermauerung dieser ersten Interpretation haben wir natürlich auch noch andere Visualisierung verwendet. Da eine D-Matrix und auch eine U-Matrix 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

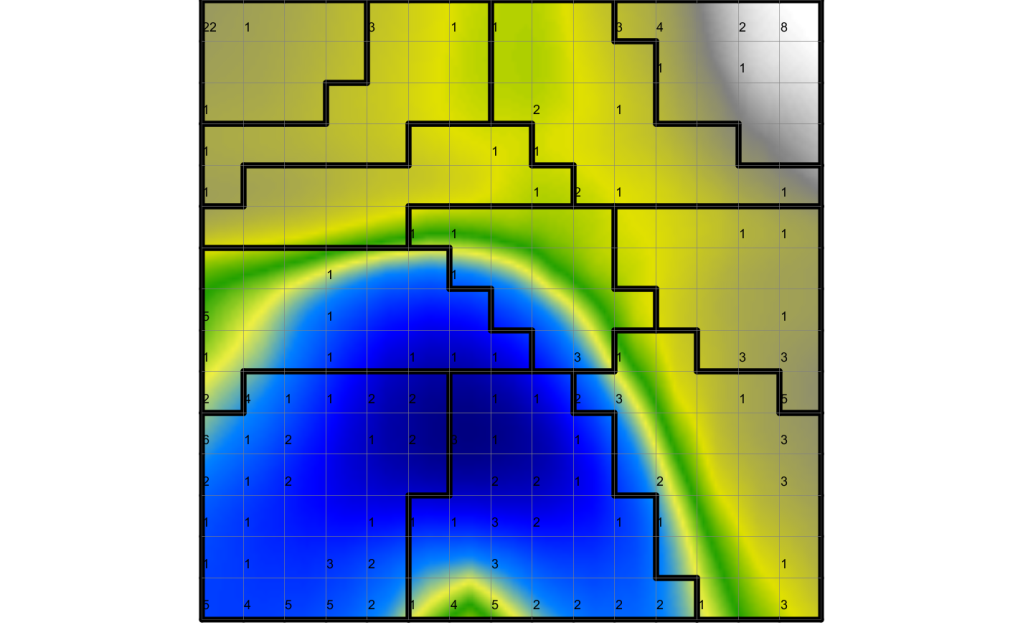
U\*-Matrix

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 „Hauptcluster“ können hier sehr gut erkannt werden. Aber auch Subcluster, auf die wir anschließend noch ein bisschen genauer eingehen, werden durch diese Visualisierung schon aufgezeigt.



Flow & Borderlines

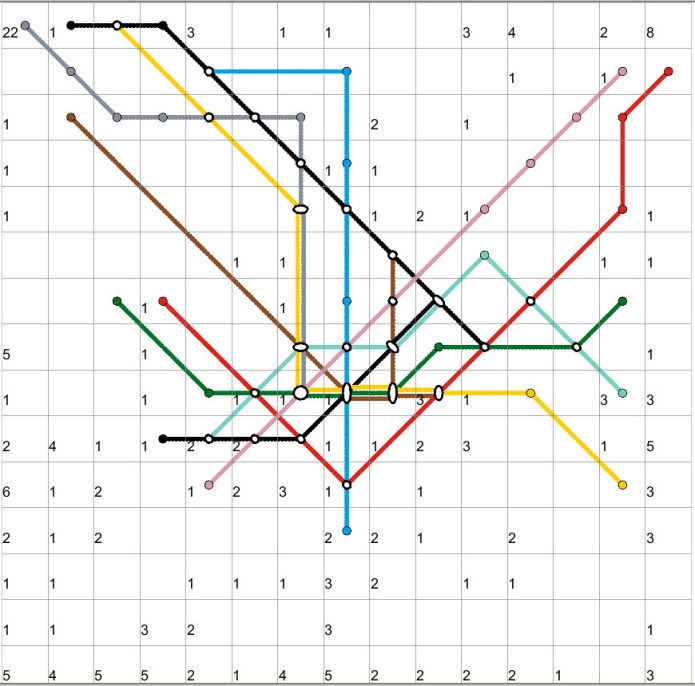
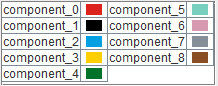
In der folgenden Abbildung haben wir versucht die Subcluster herauszuarbeiten. Wir haben die vier Cluster nochmals in elf Subcluster unterteilt. Das linke obere Cluster besteht aus 3 Sub-Cluster, das rechte obere aus 2, das darunter liegende aus 3 und das große Cluster links unter wieder aus 3 Sub-Cluster.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



Smoothed Data Histogram mit Map Sub-Clustering k-means

# Attribute

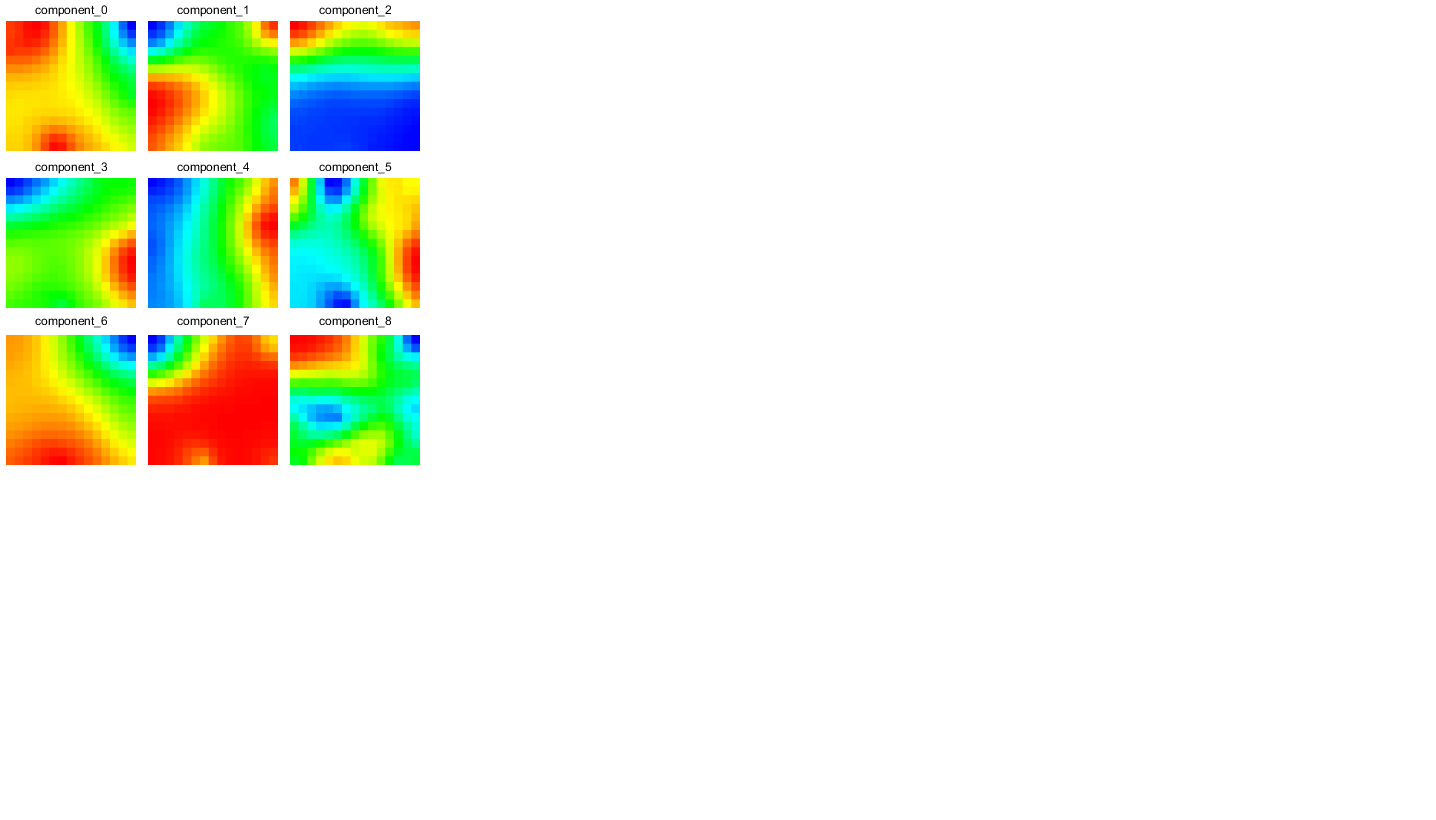
Um die verschiedenen Komponenten und deren Beziehung untereinander analysieren zu können, verwenden wir die sogenannten Metro-Maps, in denen verschiedenste Informationen abstrahiert und vereinfacht in einer Visualisierung dargestellt werden, die wesentlich besser zu interpretieren ist. Metro-Maps helfen 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

In dieser Metro-Map sehen wir 9 Component-Lines. Das bedeutet in unseren Daten haben wir 9 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.



Component-Plane der einzelnen Komponenten