Aufgabe 2.1 | Multivariate Daten[1]

In diesem Abschnitt werden Vor- und Nachteile sowie die Qualifikation und Disqualifikation verschiedener Daten für einige Visualisierungstechniken (i.e. Parallel Coordinates, Scatter Plots, Parallel Sets, Star Plot) beschrieben.

Parallel Coordinates

Die Visualisierungstechnik Parallel Coordinates (geometrische Visualisierungstechnik) verwendet pro Dimension je eine eigene vertikale Linie. Mehrere dieser sind dann parallel mit einigem Abstand anzuordnen. Konkrete Datensätze sind dann quer zu den vertikelen Linien verlaufende Linien. Diese schneiden sich an der konkreten Wertausprägung mit den vertikalen Linien. Ein Beispiel für Parallel Coordinates ist in Abbildung 1 zu sehen.

Parallel Coordinates sind besonders gut zum Erkennen von Korrelationen zwischen Variablen. Eine positive Korrelation äußert sich in horizontalen Linien, eine negative Korrelation in gekreuzten Linien zwischen zwei Dimensionen.

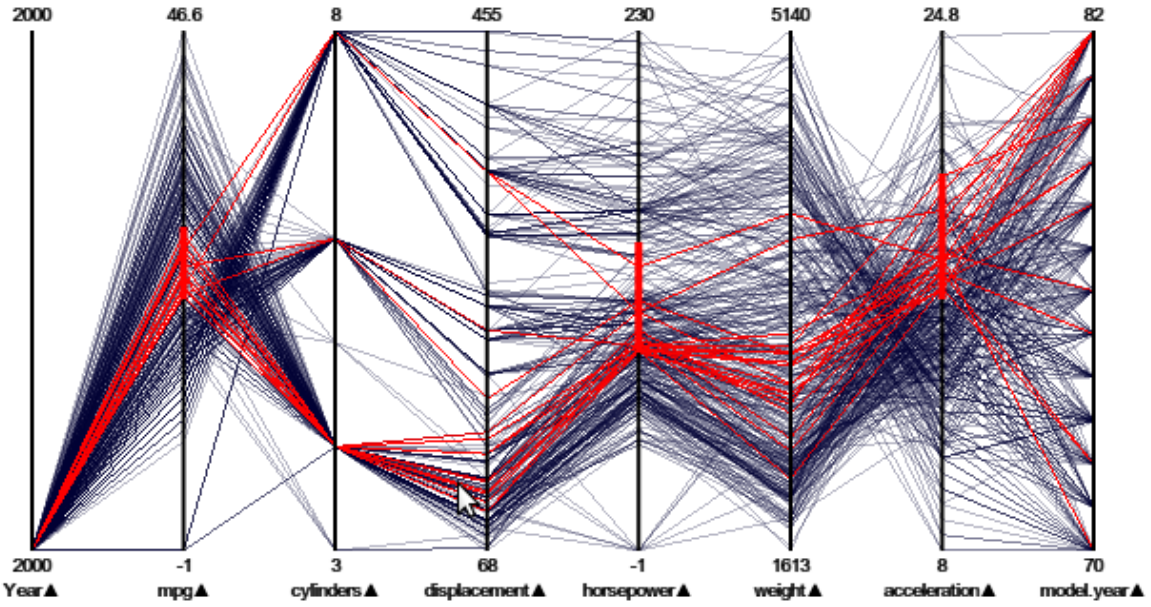

Abbildung 1: Visualisierung mit Hilfe von Parallel Coordinates

- **Vorteile / Daten die diese Methode qualifizieren:** Im Gegensatz zu Scatter Plots auch bei mehr als drei Dimensionen verwendbar. Quantitative (numerische), ordinale und nominale Daten können verwendet werden.
- **Nachteile / Daten die diese Methode disqualifizieren:** Bei zu vielen Datenpunkten sollte eingegriffen werden. Als Möglichkeiten sind hier folgende Möglichkeiten anzuden-

[1] Die Qualität der Bilder die man zum Thema Visualisierung im Internet findet lässt oft zu wünschen übrig. Weiterhin werden oft Bilder die als „png“ gespeichert werden sollten, als „jpg/jpeg“ angeboten.

ken: Brushing, das Eingrenzen von Wertebereichen, das Vertauschen von Achsen oder verschiedene Clustertechniken wie das hierarchische Clustering.

Scatter Plot

Ein Scatter Plot (geometrische Visualisierungstechnik) beschreibt Datenpunkte, eingetragen in ein 1-,2- oder 3-dimensionales Koordinatensystem — so wie man es aus der Mathematik kennt. Ein Beispiel für ein Scatter Plot ist in Abbildung 2 zu sehen.

Scatter Plots sind besonders gut zum Erkennen von Objektclustern. Cluster erkennt man optisch trivial an einer erhöhten Anzahl von Objekten in einem bestimmten Bereich im Koordinatensystem.

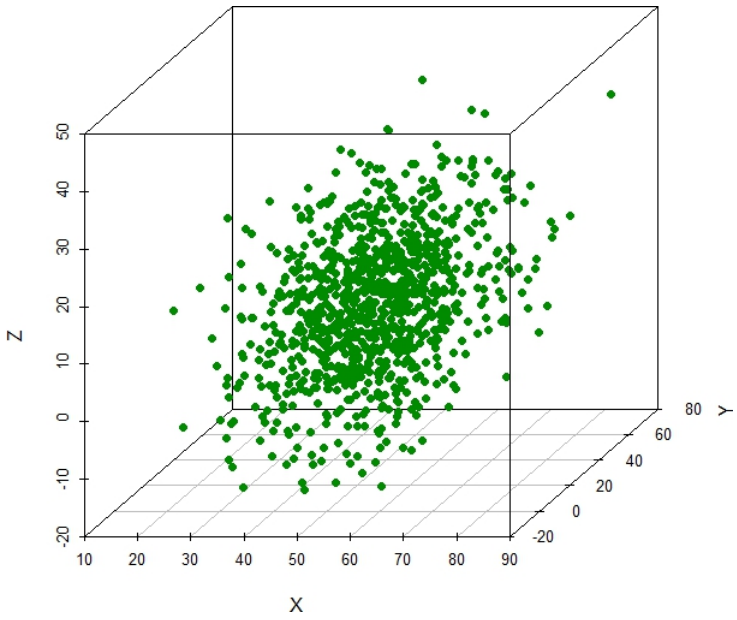

Abbildung 2: Scatter Plot mit drei Dimensionen

- **Vorteile / Daten die diese Methode qualifizieren:** Scatter Plots sind gut für eine schnelle Übersicht, zur Strukturfindung und um Zusammenhänge zu erkennen. Daten mit einer, zwei oder maximal drei Dimensionen. Große Anzahl an Datenpunkten möglich. Bei zu vielen Datenpunkten sollten mehrere Datenpunkte zu einem zusammengefasst werden (Diese Eigenschaft haben alle geometrischen Visualisierungstechniken gemeinsam). Quantitative (numerische), ordinale und nominale Daten können verwendet werden.
- **Nachteile / Daten die diese Methode disqualifizieren:** Eine hohe Dimensionalität des Datensatzes. Da Scatter Plots in einem kartesischen Koordinatensystem dargestellt werden, können nur 1 und 2 (bzw. begrenzt 3) Dimensionen dargestellt werden. Die dritte Dimension ist vom Betrachtungsmedium abhängig besser (Computer, interaktiv) oder schlechter (Papier, Computer, nicht-interaktiv) interpretierbar.

Parallel Sets

Parallel Sets ist eine geometrische Visualisierungstechnik. Eine horizontale Linie, unterteilt in verschiedene Abschnitte (je nach Ausprägungen der ersten Dimension) wird erstellt. Die verschiedenen Abschnitte gehen nun in unterschiedlicher Breite zur nächsten Dimension über. Am besten erkennt man das ganze an einem Beispiel (Abbildung 3). Man erkennt z.B., dass etwa 60% des harten Wassers („hard“) bei einer geringen Temperatur („Low“) getrunken wird. Die restlichen 40% des harten Wassers werden bei hoher Temperatur getrunken („High“).

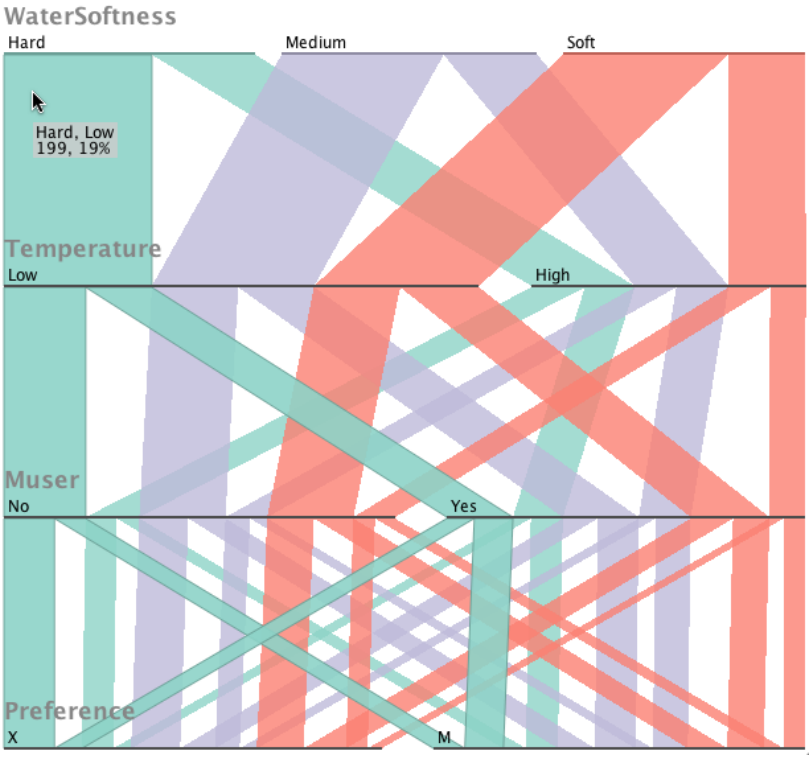

Abbildung 3: Ein Parallel Set als Beispiel

- **Vorteile / Daten die diese Methode qualifizieren:** Die Anzahl der möglichen Dimensionen ist theoretisch nicht begrenzt, allerdings verliert die Darstellung an Übersicht wenn zu viele Dimensionen verwendet werden. Auch hier ist es vom Medium abhängig wie gut eine hohe Dimensionsanzahl vom Menschen verarbeitet werden kann. Hovereffekte, die einen bestimmten Teilbaum eines Parallel Sets hervorheben könnten zum Beispiel vom Vorteil sein. Für Parallel Sets können quantitative (numerische), ordinale und nominale Daten verwendet werden.
- **Nachteile / Daten die diese Methode disqualifizieren:** Wie in den Vorteilen schon angesprochen wirkt sich eine hohe Dimensionsanzahl negativ auf die Übersichtlichkeit aus.

Star Plot

Der Star Plot (manchmal auch Star Glyph) (icon-basierte Visualisierungstechnik) besteht aus einem Mittelpunkt, einer Linie für jede Datendimension (angeordnet im Kreis) sowie schließlich jeweils eine Linie zwischen den Datendimensionslinien. Somit entsteht eine einem Spinnennetz ähnliche Struktur, ein Polygon. Ein Beispiel (hier mit drei durch Farben getrennte Datensätze) kann in Abbildung 4 betrachtet werden.

Der Star Plot ist besonders gut zum Erkennen von Ausreißern geeignet. Das rührt daher, zum einen zwischen Star Plots eine gute Vergleichbarkeit herrscht. Zum anderen stechen Ausreißer durch eine sehr steile Spitze (positiver Ausreißer) oder durch eine fast nicht sichtbare Dimensionsausprägung (negativer Ausreißer) hervor.

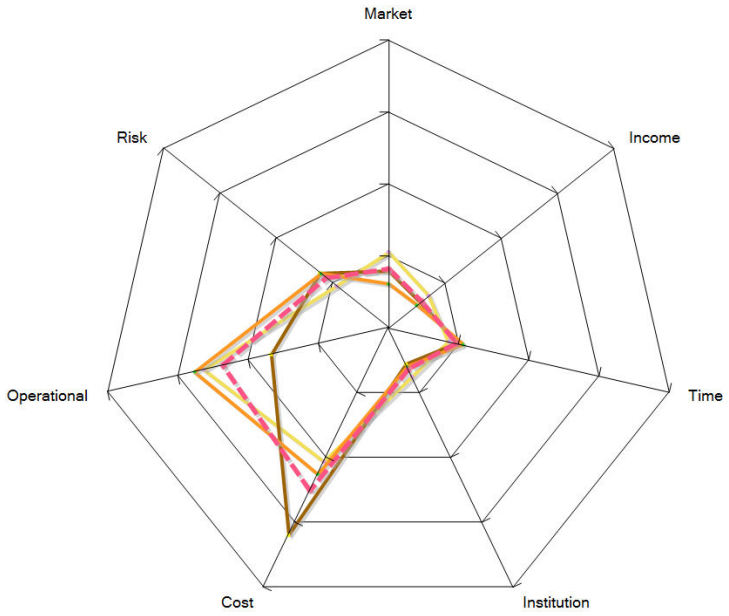

Abbildung 4: Ein Parallel Set als Beispiel

- **Vorteile / Daten die diese Methode qualifizieren:** Durch die geometrische Anordnung der Ausprägungen ist es schnell möglich, Vergleiche zwischen verschiedenen Datensätzen auszumachen.
- **Nachteile / Daten die diese Methode disqualifizieren:** Wie auch bei den geometrischen Visualisierungstechniken ist auch hier die Skalierbarkeit ein Problem. Existieren viele Dimensionen so ist das Argument der Vergleichbarkeit zwischen verschiedenen Datensätzen nicht mehr gegeben. Weiterhin sind Icon-basierte Visualisierungstechniken, also auch der Star Plot, nur nützlich wenn qualitative (und nicht quantitative) Aspekte von Datensätzen dargestellt werden sollen. Das kommt daher, dass pro Dimension effektiv nur eine sehr geringe Anzahl an Ausprägungen übersichtlich dargestellt werden kann.

Aufgabe 2.2 | RadViz

RadViz (*Radial Coordinates Visualization*) ist eine Visualisierungstechnik, mithilfe derer multivariate Daten innerhalb einer Kreisfläche dargestellt werden. Dazu werden die Variablen durch Punkte auf der Kreislinie repräsentiert. Ein Datum wird im Gegensatz zur Visualisierungstechnik der parallelen Koordinaten nur mittels eines einzigen Punktes visualisiert. Seine Lage im Kreis ergibt sich aus den jeweiligen Variablenwerten, nämlich dort wo nach dem Hooke'schen Gesetz die Summe der verschiedenen Kräfte gleich Null ist. Unter Kraft versteht man hier eine Art Anziehungskraft der Variablen-Repräsentanten, die umso stärker ist, je höher der entsprechende Variablenwert ist. So können auf den ersten Blick Erkenntnisse über eine größere Anzahl von Daten getroffen werden, ohne die einzelnen Variablenwerte vergleichen zu müssen.

Vergleich	Parallel Coordinates	RadViz
Wie leicht verständlich sind die Visualisierungstechniken?	keine Erklärung nötig	es muss zunächst verstanden werden, wie die Punkte im Kreis angeordnet werden
Wie schnell sind Erkenntnisse zu gewinnen?	ohne Hilfsmittel (Einfärben o.Ä.) relativ mühsames Finden von Mustern	sofortige Erkenntnis von Verteilungen
Für welche Datensätze (Art und Größe) eignen sie sich?	sollte verwendet werde, wenn die Einzelwerte der Dimensionen von Bedeutung sind, nur so viele Dimensionen wie Achsen in der Breite wahrnehmbar sind (schätzungsweise ~ 7)	kein Interesse an einzelnen Werten, mehr Dimensionen darstellbar durch Verteilung auf Kreislinie, eignet sich, wenn Variablen unter einem gemeinsamen Aspekt betrachtet werden können und sich im Normalfall gegenseitig beeinflussen
Können sie erweitert werden um zusätzliche Daten darzustellen?	mehr Daten: mehr Linien, höhere Grafik, Gruppieren von Datensätzen mehr Dimensionen: breitere Grafik (begrenzt)	mehr Daten: mehr Punkte, evtl. Gruppieren von Punkten mehr Dimensionen: mehr Unterteilungen auf der Kreislinie (schätzungsweise ~ 32 und mehr), Gruppieren von Variablen, dabei evtl. Verlust von Ergebnissen der Visualisierung

	Parallel Coordinates	RadViz
Welche Interaktionsmöglichkeiten (Brushing, Einfärben etc) lassen sich mit den Visualisierungen verknüpfen?	Brushing zur Hervorhebung bestimmter Datensätze, Clustering zur Reduzierung der Komplexität	Brushing nicht möglich, da keine Informationen über Einzelwerte mehr darstellbar sind, Clustering möglich, aber evtl. Verschleiern von Zusammenhängen
Was ist beim Erstellen von Visualisierungen zu bedenken?	sehr begrenzte Anzahl Dimensionen darstellbar, Auswahl der Variablen und Reihenfolge spielen eine Rolle, aber auch bei „schlechter“ Wahl sind Korrelationen zu erkennen	Auswahl der Variablen und deren Anordnung, bei ungünstiger Reihenfolge der Variablen auf der Kreislinie eventuell gar keine Erkenntnisse
Welche Vorteile bieten die Visualisierungen?	alle Einzelwerte werden sichtbar, die Grafik kann nach ersten Erkenntnissen Schritt für Schritt weiter verbessert werden	schnelle Erkenntnisse, gut einsetzbar zur Klassifizierung unbekannter Daten